How Many Labelled Trials Is Euclidean Alignment Worth? A Within-Corpus Dose-Response for EEG and EMG Transfer

Anonymous authors
Paper under double-blind review

Abstract

Euclidean Alignment (EA) and its Riemannian relatives are near-universal preprocessing in EEG and EMG decoding, and are reported as accuracy gains over an unaligned cross-subject baseline. That baseline answers the wrong question. The question a practitioner faces is not *does alignment help?* but *is borrowing donor data worth more than labelling a few trials from this user?*, and answering it requires a column the alignment literature does not report: an in-domain k-shot decoder trained on the target subject alone. Across ten public configurations and 36 within-corpus cells, the crossing budget—the number of labelled target trials at which the in-domain decoder reaches the aligned cross-subject one—ranges from 12 to over 240 total trials and depends on the decoder, so no single constant answers the question and a paper reporting a gain without this column has not reported enough. We characterise what governs the budget. It is *not* class count: a between-corpus ordering by K looked compelling and was refuted by the within-corpus dose-response we ran to test it, which finds $k^\star$ falling with K on high-SNR sEMG and flat on low-SNR EEG-MI. It is not modality either. What our data leaves standing is within-subject decodability, measurable in advance—though we show only that class count does not govern the budget, not that decodability does. Priced for the decoder a practitioner would actually deploy—tangent-space logistic regression on both arms, with the in-domain arm touching no donor data at all—the budget is **18–40 total labelled trials** on every corpus tested, a $2.2\times$ band across two modalities and $K \in \{2, 4, 6\}$. Read per user rather than per corpus, 60 total trials reaches the aligned level for 87% of sEMG and 70% of EEG-MI subject-cells, with 10% and 22% served by *no* tested budget—a ceiling set by our grid rather than an observed asymptote, and a lower bound on what the cross-subject route costs. All of this is interpretable only against a calibration of zero, which is the paper's second contribution: when the alignment reference *and* the fit are both in-domain, the gain is *identically* zero for any decoder invariant to the corresponding congruence—a two-line consequence of affine invariance, exact, falsifiable by one trial changing one label, and holding on **137 subject-fits** (119 above a pre-declared informativeness floor) on two classifier families at every regularisation tested. A factorial then separates the two knobs the literature conflates: the reference governs exactness, the fit governs magnitude.

1 Introduction

Cross-subject decoding of EMG and EEG is dominated by nuisance variation—electrode placement, impedance, anatomy—that differs per recording domain. Euclidean Alignment (He & Wu, 2020) addresses this by whitening every domain by its own reference covariance; Riemannian Alignment (Zanini et al., 2018; Rodrigues et al., 2019) uses the Fréchet mean instead; TSMNet (Kobler et al., 2022) folds the operation into a network as domain-specific batch normalisation. The literature largely treats these as competing preprocessing steps to be benchmarked.

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.

We argue this framing misses the governing structure. Alignment is a **group action**, and its usefulness is determined by the invariance group of the downstream model. Our contributions are:

1. **A column the alignment literature never imported** (§6.5). Alignment gains are reported against an unaligned cross-subject baseline; the decision they inform requires an in-domain k-shot baseline instead. We measure it on ten configurations and report the crossing budget as a *range* (6 to > 240 total labelled trials, the upper end censored) and as a per-user service curve, not as a constant.
2. **What governs the budget, and what does not** (§6.6). A within-corpus dose-response refutes class count—the moderator our own between-corpus table appeared to support—and leaves within-subject decodability, which a practitioner can measure before deciding.
3. **A calibration of zero** (§4, §6.1). MDM under the affine-invariant Riemannian metric is provably, exactly invariant to EA when reference and fit are both in-domain: $\Delta_{\mathrm{EA}} = 0.000$, $\mathbf{0/137}$ subject-fits changed. The origin of the measurement scale, not a result.
4. **A factorial localisation** (§6.2) and a ten-configuration audit with a causal intervention and six negative results correcting claims previously advanced in this line of work, five of them our own, two about this paper's own headline.

A prior hypothesis of ours—that EA helps EMG and hurts EEG—is refuted here by our own data, and we report that explicitly (§7).

What this paper is not. We propose no new alignment method, no new architecture and no new decoder, and we make no claim to improved accuracy over any published system. The contribution is a *measurement-validity* result: a quantity the field reports as a domain-adaptation gain is shown to be, in a precisely characterised regime, a property of the alignment reference instead. The appropriate comparison is therefore not against a state-of-the-art accuracy number but against the interpretation currently placed on that number, and the deliverable is a diagnostic (§8) that tells a practitioner which regime their own pipeline is in.

2 Related work

Alignment for biosignal transfer. EA (He & Wu, 2020) whitens each domain by its arithmetic-mean reference covariance; Riemannian Alignment (Zanini et al., 2018) and Riemannian Procrustes Analysis (Rodrigues et al., 2019) use the Fréchet mean, and RPA adds rescaling plus a rotation that reads target labels, placing it on our budget axis rather than in the zero-shot column (Appendix P); CORAL (Sun et al., 2016) matches second-order statistics in Euclidean space; and TSMNet (Kobler et al., 2022) learns the operation as domain-specific batch normalisation on the SPD manifold. These are typically presented as competing preprocessing steps and compared by end-task accuracy.

Riemannian geometry for EEG/EMG. Covariance decoding under the affine-invariant metric—minimum-distance-to-mean and tangent-space classifiers (Barachant et al., 2012), reviewed by Congedo et al. (2017)—is a strong classical baseline; the metric itself is due to Pennec et al. (2006). Our theory rests on two standard properties of this metric: congruence-invariance of the distance and congruence-equivariance of the Fréchet mean.

Our positioning. Prior work asks *which* alignment variant is best; we ask *when* alignment can matter at all, and answer with the invariance group of the downstream model. This yields an exact—not statistical—theorem, a factorial that localises the effect to the reference domain and the feature-space symmetry rather than the classifier family, and an audit that overturns several benchmark-level intuitions (§7). The informal prior view—alignment as removing a domain-specific transform—is here made precise and, crucially, falsifiable: it predicts an *exact* zero, which we observe on $137/137$ subject-fits.

3 Setup and data

We use public data only: nine public corpora in ten configurations (Table 1), on which the exactness certificate below rests for every leave-one-subject-out fit. The silent-speech and high-density EMG

Table 1: The public datasets. C: channels; K: classes; n: subjects contributing to the exactness certificate.

dataset	modality	n	C	K	role
csl	HD-sEMG	8	35	2	high channel count
gowda22	orofacial sEMG	12	22	3	montage subset (22 ch)
gowda3	orofacial sEMG	12	6	3	montage subset (6 ch)
shin	EEG	26	30	3	primary EEG
grabmyo	forearm sEMG	43	16	6	primary EMG
bnci2014001	EEG (motor imagery)	9	22	2	public benchmark
bnci2014004	EEG (motor imagery)	9	3	2	public benchmark
weibo2014	EEG (motor imagery)	10	60	2	public benchmark
zhou2016	EEG (motor imagery)	4	14	2	public benchmark
emguka	silent-speech sEMG	4	6	3	silent-speech check
total		**137**			10 configurations

sets are GRABMyo (Pradhan et al., 2022), an orofacial silent-speech sEMG set we label Gowda (Gowda et al., 2025; Gowda & Miller, 2024), EMG-UKA (Wand et al., 2014) and CSL-HDEMG (Diener et al., 2020); the EEG sets are Shin (Shin et al., 2018) and four motor-imagery benchmarks reached through MOABB (Jayaram & Barachant, 2018; Tangermann et al., 2012; Zhou et al., 2016; Yi et al., 2014). The protocol is leave-one-subject-out and the **subject is the unit of analysis** throughout. We use paired Wilcoxon signed-rank tests, BCa bootstrap 95% confidence intervals, and Holm–Bonferroni correction across the primary family.

The screen, and the two numbers it produces. Before any transfer claim is made on a dataset we run the same screen on every one of them, and we report *two* quantities rather than one: whether the leave-one-subject-out decoder clears chance, and what fraction of the margin the corpus's *own* within-subject decoder achieves survives when the decoder must cross subjects (Table 10). The two disagree sharply, and the disagreement sets the scale for every gain reported later. One corpus fails the within-subject check outright and is excluded, and we report why (§7). CSL-HDEMG clears *within* subject (0.557 against a chance of 0.500) but its pooled and leave-one-subject-out decoders both sit at chance (0.513, 0.513): its subjects are mutually inconsistent, so it contributes to the reference-size and rank-reduction analyses and we draw no transfer conclusion from it. Shin clears the screen unambiguously in absolute terms (0.397 against a chance of 0.333; $t \approx 4.5$ over 26 subjects) and yet retains only 12% of the 0.525 margin its own within-subject decoder reaches—the second-lowest of our configurations.

The Gowda 6- and 22-channel rows are the same 12 subjects under a montage subset we selected; they are *not* independent corpora, and Appendix D records the selection, its dependence on three of these twelve subjects, and the direction of the resulting bias.

4 Theory

EA whitens trials of domain s by $C \mapsto W_s C W_s^\top$, with $W_s = \bar{R}_s^{-1/2}$ the inverse square root of that domain's reference covariance. Crucially, W_s depends on the *domain*, not on the trial or the label.

Two remarks before the facts. First, "the" Fréchet mean is well defined here because SPD under AIRM is a Hadamard manifold—complete, simply connected, of nonpositive curvature—on which the variance functional $M \mapsto \sum_i d(C_i, M)^2$ is geodesically strongly convex and therefore has a unique minimiser (Pennec et al., 2006). Every appeal below to *the* mean rests on that convexity. Second, the property doing the work in what follows is that a congruence is an *isometry*, which is strictly stronger than preserving geodesic convexity: a shear of $\mathbb{R}^n$ preserves convexity and changes every distance. Appendix B bounds that class, and retracts an earlier claim that we had made it exact.

We use two standard facts. AIRM is congruence-invariant: $d(AC_1A^\top, AC_2A^\top) = d(C_1, C_2)$; and the Fréchet mean is congruence-equivariant: $\operatorname{mean}\{AC_iA^\top\} = A \operatorname{mean}\{C_i\} A^\top$, which follows

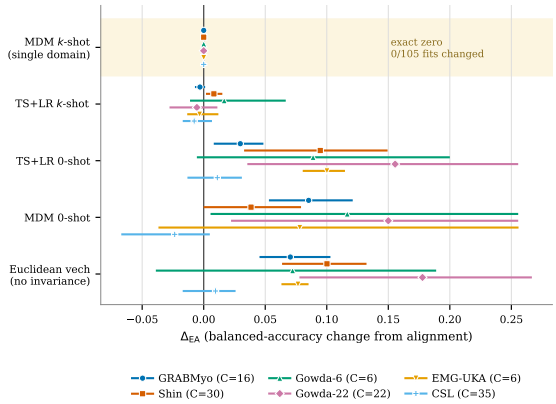

Figure 1: The invariance ladder. Δ_{EA} per rung, one line per dataset, ordered by how much of the per-domain congruence the model's invariance group fails to absorb. Every dataset is exactly 0 at MDM k-shot (single-domain; $0/137$ subject-fits changed) and large on every broken-invariance rung. The ordering *within* the broken rungs is not monotone—we claim the dichotomy and the approximate middle rung, not a total order.

immediately from the first since the mean minimises a sum of *squared* distances (Pennec et al., 2006; Congedo et al., 2017).

Proposition 1. *MDM under AIRM is exactly EA-invariant* whenever *the class prototypes and the test trials come from the same domain.*

Single-domain. Prototypes are Fréchet means of subject s's trials, so EA maps $P_c \mapsto W_s P_c W_s^\top$; the test trial maps by the *same* W_s; every distance $d(P_c, X)$ is unchanged; the $\arg\min$ over c is bit-identical. Hence $\Delta_{\mathrm{EA}} = 0$ exactly. *Cross-domain.* P_c is built from donors $t \neq s$, so it maps by W_t while the test trial maps by W_s; no single congruence acts on both arguments, and congruence-invariance offers no protection. A full proof, including why tangent-space logistic regression is only *approximately* invariant—the reference may live in another domain, and even with a shared reference the tangent coordinates rotate by a fixed orthogonal map to which L_2-regularised linear classifiers are equivariant but per-feature scalers are not—is given in Appendix A.

The general principle: **alignment is useful exactly to the extent it removes a transform the model's invariance group cannot absorb.**

5 Experimental design

We measure a five-rung *invariance ladder* ordered by how much of the per-domain congruence the model fails to absorb: MDM k-shot (prototypes from the held-out subject's own shots—single-domain), TS+LR k-shot, TS+LR zero-shot, MDM zero-shot, and a Euclidean decoder on $\mathrm{vech}(C)$ (no invariance). Each rung is run $\pm$EA. A *factorial* then crosses $\{$fit data: own vs. others$\}$ $\times$ $\{$tangent reference: own vs. others$\}$ $\times$ $\{$StandardScaler: on vs. off$\}$ on the same covariances, isolating the cause of the exact-zero. Exactness is defined as **zero subjects changed**, not merely a zero mean—the two come apart, and the subject count is the honest statistic.

6 Results

6.1 Calibrating zero: exact invariance, and where it breaks

On GRABMyo the MDM k-shot rung is exactly EA-invariant while every broken-invariance rung is large (Table 2). The signature replicates on all datasets: Δ_{EA} is exactly 0.000 with $\mathbf{0/137}$ subject-fits changed (Figure 1; per-dataset in Appendix O, Tables 8 and 9).

Table 2: Invariance ladder on grabmyo ($n = 43$). The exactness is the point: zero changed predictions, not a small effect. TS+LR k-shot is ≈ 0 but not exactly, as predicted for an approximately-invariant pipeline.

rung	raw	+EA	Δ_{EA}	95% CI	p	subj. changed
MDM k-shot (single-domain)	0.771	0.771	$+\mathbf{0.000}$	$[+0.000, +0.000]$	1.000	$\mathbf{0/43}$
TS+LR k-shot	0.933	0.930	-0.003	$[-0.007, +0.001]$	0.117	37/43
TS+LR zero-shot	0.775	0.805	$+0.030$	$[+0.008, +0.049]$	$4.5\mathrm{e}{-4}$	42/43
MDM zero-shot	0.586	0.671	$+0.085$	$[+0.053, +0.121]$	$1.3\mathrm{e}{-5}$	43/43
Euclidean vech	0.761	0.831	$+0.070$	$[+0.045, +0.103]$	$4.8\mathrm{e}{-6}$	43/43

How many of those 137 are informative. A zero is evidence only where the classifier classifies. On a pre-declared floor—informative if raw accuracy clears chance by 0.05—the rung is at chance on Weibo2014 and marginal on CSL, the two highest-dimensional cells, so the certificate rests on $\mathbf{119/137}$ informative fits and we quote both (Appendix O, Table 9).

A second classifier family, needing strictly less. MDM's exactness uses congruence-invariance of the distance *and* congruence-equivariance of the Fréchet mean, so a reader may suspect it is a property of nearest-mean classification. It is not: k-NN under AIRM forms *no mean at all* and needs only the first fact. On the same single-domain protocol, $\mathbf{0/1710}$ predictions changed on BCI IV-2a and $\mathbf{0/7310}$ on GRABMyo—$9{,}020$ predictions. The certificate is a property of the metric, not the mean; Appendix B states what it does and does not cover.

Reading this as a zero, not a result. The proposition is two lines of standard Riemannian geometry and $0/137$ is what they predict. It earns its space by fixing the origin of the scale on which published Δ_{EA} is read: here the quantity *cannot* be non-zero, so a non-zero value reports a pipeline's invariance group. The rest of the paper measures what happens once that regime is left.

6.2 The exact-zero tracks the reference, not the classifier

The ladder confounds the classifier family with where the reference point lives. The factorial decouples them on the same covariances (Figure 2). Exactly two cells are exact—MDM with in-domain prototypes, and TS+LR with an in-domain reference *and no scaler*—and both are exact on every subject of all three datasets ($\mathbf{0/81}$), while every `fit-others` cell is large. Two structural predictions of §4 hold. **(P1)** Holding the classifier at TS+LR · fit-own · no-scaler and moving only the reference from the target to the donor domain turns the exact zero into a non-zero effect on all three sets ($|\Delta_{\mathrm{EA}}|$ for ref=others $>$ ref=own everywhere). **(P2)** Adding a StandardScaler—which rescales each tangent coordinate and so does not commute with the residual rotation EA leaves behind—changes $31/43$ subjects on GRABMyo and $26/26$ on Shin, versus 0 without it; on Gowda ($C = 6$) the scaler flips no labels either way, so that set is non-informative for P2 rather than a counter-example.

This reconciles the two "TS+LR k-shot" numbers. The ladder linearises the tangent space at the pooled training-subject (donor) mean, so its k-shot rung ($37/43$ changed) is the factorial's own-fit/*donor*-reference/scaler cell ($38/43$); moving that reference in-domain (own-reference cell) reduces the flips to $31/43$, and removing the scaler as well takes them to $0/43$. Neither move alone suffices: the exact zero needs the in-domain reference *and* the absence of a per-feature scaler, which is precisely the conjunction §4 predicts. The two counts are two different cells, not two estimates of one; we verified per-subject identity of the ladder rung and the matching factorial cell on synthetic covariances.

6.3 Alignment helps EEG as much as EMG

Alignment helps both modalities; a prior hypothesis of ours that it helps EMG and hurts EEG is refuted by our own data. Three of ten configurations survive Holm–Bonferroni over the declared family (GRABMyo $+0.071$, Shin $+0.101$, Gowda-22 $+0.167$), all ten point estimates are positive

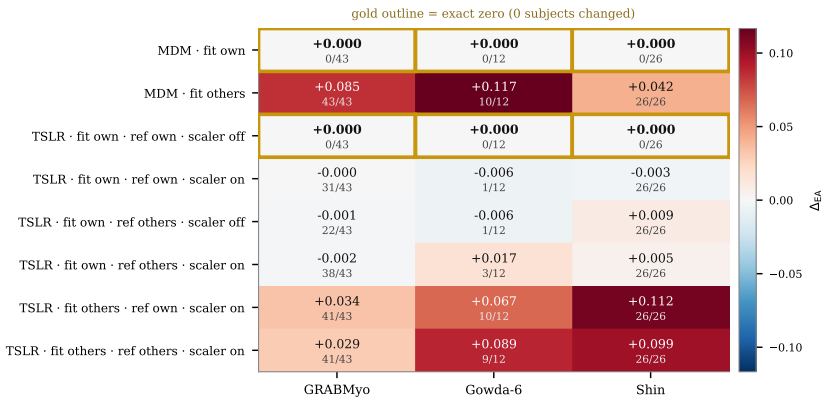

Figure 2: Reference-point factorial. Colour encodes Δ_{EA} (magnitude of the alignment effect); the annotation is the honest exactness statistic (subjects whose prediction changed); a gold outline marks cells with zero changed. The two exact cells—MDM fit-own and TS+LR own-reference/no-scaler—are exact across all three datasets; Gowda's extra outlined cells are its P2 saturation. The exactness is a property of *reference domain* × *feature-space symmetry*, not of the classifier family.

and eight of ten intervals exclude zero. Appendix I gives the per-corpus table, the declared family and the correction.

6.4 A REGULARISATION CONSTANT THAT CHANGES THE CONCLUSION

Every covariance pipeline in this literature applies a shrinkage $C \mapsto (1-\alpha)C + \alpha\,\mathrm{tr}(C)/p\,I$, and we could find no prior work that reports sweeping it. It is not a nuisance parameter: on Shin, moving α from 0 to 0.1 takes Δ_{EA} from $+0.060$ to -0.0001—the effect is eliminated and its sign reverses; no fold-change describes a quantity that crosses zero—and costs the nearest-mean k-shot arm 26 accuracy points; on GRABMyo the cost is 9.3 points. The exactness certificate is untouched ($\Delta_{\mathrm{EA}} = 0$ at every α, $25/25$ cells), because it is algebraic. Every *magnitude* in this literature, however, is quoted at an α that is usually not stated. We report at $\alpha = 0$ throughout and stamp the value into every result record.

A causal displacement intervention. Injecting a synthetic electrode displacement and measuring recovery gives the one genuinely causal test here: alignment recovers decoding under displacements up to $\delta = 0.7$ and fails beyond (Appendix F). This supports the nuisance-removal account in the regime where it is testable, and is reported in full rather than summarised because it cuts against the paper's framing.

6.5 THE EXCHANGE RATE, AND THE VARIABLE THAT DECIDES IT

§6.2 shows the magnitude lives in the fit boundary. That leaves the question a practitioner asks: *align to donors, or collect labelled trials from this user?* Answering it needs a comparison of *levels* between two rungs—aligned cross-subject MDM against in-domain k-shot MDM with none—and to our knowledge no work reporting an alignment gain has run it. Nor had we: two earlier versions of this paper asserted opposite answers—five trials, then "tens of trials"—each generalised from corpora that shared a property we had not thought to record.

Design. The ladder's calibration split shrinks the test set as k grows, so its levels are not comparable across k. We reserve one test block per subject—half of each class, drawn before any k—and draw calibration trials from the disjoint remainder, averaging three to five draws per k. Test block, donor pool and subjects are identical at every k; only the calibration trials vary, and the per-subject trial cap is declared with each result because it changes the answer.

Table 3: **The exchange rate.** Aligned cross-subject MDM (+align) against in-domain k-shot MDM with *no* alignment, on a test block reserved before k is chosen and held identical across every k. $k^\star$ is the first grid point at which the unaligned in-domain decoder reaches the aligned level: a crossing of two *means*, whose wide intervals and per-subject crossings Appendix Q gives. †: excluded from pricing (gain interval spans zero, or the in-domain arm is itself at chance). $\leq$: the pool admits only $k = 5$, so the crossing is bounded, not located; note 5 is the grid floor throughout. ‡: this pool alone admits $k > 60$; swept to $k = 120$ the in-domain curve keeps rising (0.683, 0.686, 0.685) to 0.008 short of aligned, and 4 of 9 subjects do cross their own aligned level, so $k^\star > 120$ per class (> 240 total) is censored, not unbounded. The apparent ordering of $k^\star$ by class count is refuted in §6.6.

				cross-subject			in-domain k-shot, unaligned					
dataset	mod.	K	n	x-subj	+align	gain (95% CI)	5	10	20	40	60	$k^\star$
GRABMyo	sEMG	6	43	0.585	0.671	+0.087 [+0.05, +0.12]	0.771	0.813	0.843	–	–	5
EMG-UKA†	sEMG	3	4	0.438	0.453	+0.016 [−0.04, +0.08]	0.591	0.611	0.640	0.654	0.660	5
Gowda-22	sEMG	3	12	0.533	0.672	+0.139 [+0.06, +0.23]	0.867	–	–	–	–	≤ 5
Gowda-6	sEMG	3	12	0.478	0.606	+0.128 [+0.03, +0.23]	0.744	–	–	–	–	≤ 5
Shin†	EEG	3	26	0.407	0.441	+0.034 [−0.00, +0.07]	0.688	0.733	0.765	0.787	0.798	5
BCI IV-2a	EEG	2	9	0.615	0.731	+0.116 [+0.06, +0.17]	0.640	0.678	0.695	0.725	0.745	60
BCI IV-2b	EEG	2	9	0.672	0.694	+0.022 [+0.00, +0.04]	0.593	0.628	0.659	0.675	0.675	$> 120^\ddagger$
CSL†	sEMG	2	8	0.564	0.547	−0.016 [−0.04, +0.00]	0.527	0.543	0.560	0.575	0.577	20
Weibo2014†	EEG	2	10	0.583	0.679	+0.096 [+0.06, +0.14]	0.522	0.527	0.528	–	–	>20
Zhou2016†	EEG	2	4	0.703	0.730	+0.027 [−0.03, +0.09]	0.607	0.656	0.693	0.713	0.726	>60

Result, in per-class units. On all five multi-class corpora five labelled trials per class match or beat the aligned cross-subject pipeline, and on none of the five binary corpora does $k = 5$: the $5/5$ split is a statement about $k = 5$, and located binary crossings run from $k = 20$ (CSL, negative gain) to > 120 (BCI IV-2b). We first read this as a class-count effect. It is not, and §6.6 is the control that says so.

Result, in total labels, and per user. Denominated in *total* labelled trials rather than trials per class, the measurement stops depending on the moderator §6.6 refutes. We report it as a *service curve* rather than a cell fraction (Table 7), because a cell fraction is a mean over subjects and a mean is not a user. 60 total labelled trials reach the aligned cross-subject level for **87**% of GRABMyo subject-cells and **70**% of BCI IV-2a's, and **10**% and **22**% respectively are served by *no* tested budget. Table 5 gives the cell-level form ($30/30$ evaluable cells at T=60); it overstates what a deployment gets, and the two cells separating it from $34/36$ clear by $+0.002$ and $+0.0008$ without intervals.

The budget is per decoder family. A budget priced with a weak decoder on both arms is not evidence about a deployment that uses a strong one, so we swept k with nearest-mean and tangent-space logistic regression on *both* arms, the in-domain arm touching no donor data at all—not the classifier and not the tangent base point, which is the Fréchet mean of the subject's own kK calibration trials.

Read the diagonal: a practitioner chooses whether to label trials for the decoder they intend to deploy, so in-domain and aligned belong to the same family. For tangent-space logistic regression the crossing budget is **18**, **20** and **40** total labelled trials on GRABMyo, four-class BCI IV-2a and binary BCI IV-2a—a $2.2\times$ band across two modalities, two SNR regimes and K from 2 to 6. That is the tightest form in which we can state the result, and it is the form we report.

A decoder upgrade does not uniformly help the in-domain route. The budget falls $40 \rightarrow 20$ on four-class IV-2a and becomes finite on binary IV-2a where nearest-mean never crosses, but *rises* $12 \rightarrow 18$ on GRABMyo, where the aligned tangent-space bar is far higher (0.826 against 0.671) and the better decoder lifts both arms. What is defensible is that the budget stays within an order of magnitude under the upgrade, not that stronger decoders favour labelling; an earlier version asserted the latter from two cells and we withdraw it, as we withdraw the claim that this comparison's

Table 4: **The budget is per decoder family.** Aligned cross-subject against an in-domain k-shot decoder that touches *no* donor data—not the classifier, not the tangent base point (the Fréchet mean of the subject's own kK calibration trials). The diagonal, where in-domain and aligned are the same family, is the comparison a practitioner faces. For tangent-space logistic regression the crossing budget is 18–40 total labelled trials on every corpus tested. Value after each decoder name is that family's aligned level—the bar the in-domain arm must clear.

					in-domain k-shot, no donor data					crossing	
dataset	mod.	K	n	decoder	k=2	k=5	k=10	k=20	k=40	$k^\star$	total
GRABMyo	sEMG	6	43	MDM ($\rightarrow$0.671)	0.698	0.771	0.813	0.843	–	**2**	**12**
				TS+LR ($\rightarrow$0.826)	0.764	0.899	0.955	0.979	–	**3**	**18**
BCI IV-2a	EEG-MI	4	9	MDM ($\rightarrow$0.479)	0.415	0.465	0.504	0.531	0.556	**10**	**40**
				TS+LR ($\rightarrow$0.503)	0.430	0.523	0.590	0.645	0.702	**5**	**20**
BCI IV-2a	EEG-MI	2	9	MDM ($\rightarrow$0.731)	0.609	0.645	0.662	0.688	0.719	never	–
				TS+LR ($\rightarrow$0.735)	0.619	0.668	0.697	0.739	0.765	**20**	**40**

Table 5: **The within-corpus dose-response, and the retraction it forces.** Class subsets are sampled with a fixed seed before any result is seen; every row is a median over subsets, with subjects, channels and preprocessing held fixed and only K varying. *Per-class* $k^\star$ falls with K on GRABMyo and is *flat* on BCI IV-2a, so the between-corpus class-count ordering of Table 3 does not reproduce within a low-SNR corpus. The total at the crossing rises with K on both. 'clears at T' is the fraction of cells in which a fixed *total* budget of T labelled target trials beats the aligned cross-subject decoder.

corpus	K	cells	$k^\star$	total $= k^\star K$	in-dom. SNR	T=30	T=60	T=120
GRABMyo (sEMG, n=43)	2	6	3	6	+0.81	6/6	–	–
	3	6	2	6	+0.76	6/6	6/6	–
	4	6	2	8	+0.79	6/6	6/6	–
	5	6	2	10	+0.74	6/6	6/6	–
	6	1	2	12	+0.73	1/1	1/1	1/1
BCI IV-2a (EEG-MI, n=9)	2	6	10	20	+0.34	5/6	6/6	6/6
	3	4	10	30	+0.31	3/4	4/4	4/4
	4	1	10	40	+0.29	1/1	1/1	1/1
all cells	2–6	36				**34/36**	**30/30**	**12/12**

direction was undetermined while the relevant numbers sat in Table 2. The only prohibitive pairing is the mismatched one—nearest-mean in-domain against tangent-space aligned, 120 labels on GRABMyo—which no deployment faces.

6.6 The within-corpus control, and what it refutes

A between-corpus ordering is not a mechanism: corpora differ in subjects, channels, paradigm and SNR as well as in K, and among the five corpora carrying a priced gain, K is perfectly collinear with modality. We therefore varied K *inside* two corpora—GRABMyo (K up to 6, high-SNR sEMG, $n = 43$) and BCI IV-2a re-ingested at its native four classes (low-SNR EEG-MI, $n = 9$)—holding subjects, channels, montage, preprocessing, solver and seed fixed, sampling class subsets with a declared seed before any result was seen, and marginalising $k^\star$ over those subsets. We pre-registered the reading: a slope inside the low-SNR EEG-MI corpus identifies K; a slope confined to high-SNR sEMG means the between-corpus pattern was confounded.

The class-count reading is refuted. Per-class $k^\star$ falls with K on GRABMyo ($3 \rightarrow 2$, partial rank correlation with within-subject decodability controlled, $\rho = -0.55$) and is *flat* on BCI IV-2a—

median 10 at $K = 2$, 3 and 4 alike, with $\rho = +0.52$. The ordering does not reproduce where it would have to. We withdraw class count as an explanation.

The per-class denominator was doing part of the work: on GRABMyo $k^\star$ falls from 3 to 2 while the *total* at the crossing rises from 6 to 12, so only the budget is monotone on both corpora. With two corpora we cannot establish that within-subject decodability governs $k^\star$ either; we can say that class count does not.

What we do not claim. Four configurations are excluded from pricing and marked in Table 3: EMG-UKA, Shin and Zhou2016 have gain intervals spanning zero and CSL's is negative, so they have no gain to price; Weibo2014's in-domain arm is itself at chance (0.522–0.528 against 0.500, the same cell the certificate flags as uninformative), so its $k^\star > 20$ reports a broken baseline rather than a valuable alignment. That leaves five priced corpora, three multi-class and two binary—by the standard of §6.7 a direction, not a law. The donor pool is large and the k-shot arm small, so this is the deployment question, not a controlled comparison of estimators. Three limits bound what the decoder-family result in Table 4 supports. It rests on two corpora, not ten. Its margins carry no confidence intervals, so a crossing that clears the aligned bar by a few thousandths—and two do—is a point estimate and nothing more. And the in-domain arm is deliberately the austere version, using none of the donor pool even for its tangent base point, which makes it a lower bound on what a practitioner with any donor data could achieve.

6.7 A REFERENCE-SIZE REGIME WHERE ALIGNMENT IS HARMFUL

The alignment reference itself needs data. On GRABMyo ($n = 43$) the gain interval lies *below* zero at 2, 4 and 8 reference trials and first clears zero at 16; the harmful regime replicates on three of nine corpora, so it is real and not universal. The mechanism is the familiar one—a reference estimated from too few trials injects more estimation noise than it removes nuisance—but neither the original EA proposal (He & Wu, 2020) nor the recent tutorial treatment (Wu, 2025) states a minimum, and while sensitivity to reference size is not new—the calibration-reduction literature has long traded sample size against adaptation quality (Lotte, 2015; Wu, 2017)—the declared threshold and the negative sign below it are. Appendix H gives the per-corpus curves.

6.8 THE SESSION-VERSUS-USER ORDERING DOES NOT GENERALISE

Alignment helps at least as much across sessions as across users ($+0.045$, 95% CI 0.012–0.076 on GRABMyo), contrary to the inter-subject framing that motivates it. The ordering replicates on one of five corpora, and once training-set size is matched the residual difference spans zero, so what we establish is the ordering and not its cause. Appendix J gives the declared family, the correction, the matched-n control and the four corpora on which it does not replicate.

7 NEGATIVE RESULTS

Six hypotheses advanced in this line of work, five of them ours, are refuted or withdrawn here, including both of our own earlier exchange-rate claims—five trials, then "tens of trials." Appendix N states each, the evidence against it, and what survives.

8 DISCUSSION

What this says about users, not pipelines. Alignment removes most of the between-subject variance—29–68% of the total, driven to 10^{-7} of its value (Appendix C)—yet these corpora keep only 11–81% of their within-subject margin across subjects (Table 10). What alignment removes is therefore not what limits transfer. A congruence is a sensor-level mixing, so what alignment absorbs is instrumental and the residue that defeats transfer is not.

Add the column. A gain measured against an unaligned cross-subject baseline does not answer the practitioner's question; that comparison is against an in-domain k-shot decoder, denominated in total labelled trials and read per user. The budget spans 12 to over 240 trials, so we offer no constant (Appendices L, M).

AI USE STATEMENT

An AI assistant was used for research execution, for drafting, and for editing. Specifically: it implemented and debugged the analysis harness that produces the reported statistics; it identified several methodological errors in earlier versions of that harness, including a shrinkage constant hard-coded at two pipeline sites that pinned three reported cells to an undeclared value, and several verdicts drawn from point estimates without confidence intervals; it proposed the provenance and interval-enforcement layers now shipped with the code; and it drafted the abstract and assisted in editing the manuscript. All experimental design decisions, all interpretation of results, and every judgement about what a result does and does not support are the author's. Synthetic data were used only as test fixtures inside the harness's self-tests; no reported result derives from synthetic data. The released code includes those self-tests and the per-cell provenance records, so every number in this paper can be regenerated and audited independently.

REPRODUCIBILITY STATEMENT

All code, data splits, seeds, and per-subject results are released in the supplementary material accompanying this submission. All datasets are public (Table 1); the subject is the unit of analysis throughout, with paired Wilcoxon tests, BCa bootstrap CIs, and Holm–Bonferroni correction. Every script carries a `--selftest` that validates its logic on synthetic data with known ground truth, and every number in the paper is traceable to a machine-readable results log that records its own experimental conditions—shrinkage α, featuriser, protocol arm, reference source, and a hash of the harness that produced it. The aggregator refuses to pool records whose conditions differ, and **downgrades any verdict asserted without a confidence interval to UNSUPPORTED**; the surviving exemptions are listed with written reasons in the released `alpha_allowlist.json`. Several claims in earlier versions of this work were retracted by that mechanism rather than by us, and we have said so at each point. The proposition of §4 is proved in full in Appendix A.

ETHICS STATEMENT

This work uses only previously-released, de-identified public datasets of EEG and surface EMG, under their original licenses; we collected no new human-subject data. Two qualifications are owed. Within-subject accuracies here are not negligible—0.858 on an orofacial silent-speech sEMG corpus and 0.939 on forearm sEMG (Table 10)—and it is the *transfer* results, which is what this paper measures, that remain far from a deployable capability. Separately, the mechanism we study implies that a per-subject reference covariance is individually distinctive enough that whitening by it changes every prediction; such a reference is therefore subject-identifying and is itself a data-protection object, not merely a preprocessing constant.

REFERENCES

Alexandre Barachant, Stéphane Bonnet, Marco Congedo, and Christian Jutten. Multiclass brain–computer interface classification by Riemannian geometry. *IEEE Transactions on Biomedical Engineering*, 59(4):920–928, 2012.

Marco Congedo, Alexandre Barachant, and Rajendra Bhatia. Riemannian geometry for EEG-based brain–computer interfaces: A primer and a review. *Brain-Computer Interfaces*, 4(3):155–174, 2017.

Lorenz Diener, Mehrdad Roustay Vishkasougheh, and Tanja Schultz. CSL-EMG_Array: An open access corpus for EMG-to-speech conversion. In *Proc. Interspeech*, pp. 3745–3749, 2020. doi: 10.21437/Interspeech.2020-2859.

Harshavardhana T. Gowda and Lee M. Miller. EMG – speech decoding, 2024. URL `https://osf.io/ym5jd/`. Dataset, CC BY 4.0.

Harshavardhana T. Gowda, Zachary D. McNaughton, and Lee M. Miller. Geometry of orofacial neuromuscular signals: speech articulation decoding using surface electromyography. *Journal of Neural Engineering*, 22(4):046001, 2025. doi: 10.1088/1741-2552/ade7af.

He He and Dongrui Wu. Transfer learning for brain–computer interfaces: A Euclidean space data alignment approach. *IEEE Transactions on Biomedical Engineering*, 67(2):399–410, 2020.

Vinay Jayaram and Alexandre Barachant. MOABB: trustworthy algorithm benchmarking for BCIs. *Journal of Neural Engineering*, 15(6):066011, 2018. doi: 10.1088/1741-2552/aadea0.

Reinmar Kobler, Jun-ichiro Hirayama, Qibin Zhao, and Motoaki Kawanabe. SPD domain-specific batch normalization to crack interpretable unsupervised domain adaptation in EEG. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Fabien Lotte. Signal processing approaches to minimize or suppress calibration time in oscillatory activity-based brain–computer interfaces. *Proceedings of the IEEE*, 103(6):871–890, 2015. doi: 10.1109/JPROC.2015.2404941.

Xavier Pennec, Pierre Fillard, and Nicholas Ayache. A Riemannian framework for tensor computing. *International Journal of Computer Vision*, 66(1):41–66, 2006.

Ashirbad Pradhan, Jiayuan He, and Ning Jiang. Multi-day dataset of forearm and wrist electromyogram for hand gesture recognition and biometrics. *Scientific Data*, 9(1):733, 2022. doi: 10.1038/s41597-022-01836-y.

Pedro Luiz Coelho Rodrigues, Christian Jutten, and Marco Congedo. Riemannian Procrustes analysis: Transfer learning for brain–computer interfaces. *IEEE Transactions on Biomedical Engineering*, 66(8):2390–2401, 2019.

Jaeyoung Shin, Alexander von Lühmann, Do-Won Kim, Jan Mehnert, Han-Jeong Hwang, and Klaus-Robert Müller. Simultaneous acquisition of EEG and NIRS during cognitive tasks for an open access dataset. *Scientific Data*, 5:180003, 2018. doi: 10.1038/sdata.2018.3.

Karl-Theodor Sturm. Probability measures on metric spaces of nonpositive curvature. In *Heat Kernels and Analysis on Manifolds, Graphs, and Metric Spaces*, volume 338 of *Contemporary Mathematics*, pp. 357–390. American Mathematical Society, 2003.

Baochen Sun, Jiashi Feng, and Kate Saenko. Return of frustratingly easy domain adaptation. In *AAAI Conference on Artificial Intelligence*, 2016.

Michael Tangermann, Klaus-Robert Müller, Ad Aertsen, Niels Birbaumer, Christoph Braun, Clemens Brunner, Robert Leeb, Carsten Mehring, Kai J. Miller, Gernot R. Müller-Putz, Guido Nolte, Gert Pfurtscheller, Hubert Preissl, Gerwin Schalk, Alois Schlögl, Carmen Vidaurre, Stephan Waldert, and Benjamin Blankertz. Review of the BCI competition IV. *Frontiers in Neuroscience*, 6:55, 2012. doi: 10.3389/fnins.2012.00055.

Michael Wand, Matthias Janke, and Tanja Schultz. The EMG-UKA corpus for electromyographic speech processing. In *Proc. Interspeech*, pp. 1593–1597, 2014. doi: 10.21437/Interspeech.2014-379.

Dongrui Wu. Online and offline domain adaptation for reducing BCI calibration effort. *IEEE Transactions on Human-Machine Systems*, 47(4):550–563, 2017. doi: 10.1109/THMS.2016.2608931.

Dongrui Wu. Revisiting Euclidean alignment for transfer learning in EEG-based brain–computer interfaces. *Journal of Neural Engineering*, 22(3):031005, 2025. doi: 10.1088/1741-2552/addd49.

Weibo Yi, Shuang Qiu, Kun Wang, Hongzhi Qi, Lixin Zhang, Peng Zhou, Feng He, and Dong Ming. Evaluation of EEG oscillatory patterns and cognitive process during simple and compound limb motor imagery. *PLOS ONE*, 9(12):e114853, 2014. doi: 10.1371/journal.pone.0114853.

Paolo Zanini, Marco Congedo, Christian Jutten, Salem Said, and Yannick Berthoumieu. Transfer learning: A Riemannian geometry framework with applications to brain–computer interfaces. *IEEE Transactions on Biomedical Engineering*, 65(5):1107–1116, 2018.

Ran Zhao, Yanru Bai, Shuming Zhang, Jianing Zhu, Hongxing Liu, and Guangjian Ni. An open dataset of multidimensional signals based on different speech patterns in pragmatic Mandarin. *Scientific Data*, 12(1):1934, 2025a. doi: 10.1038/s41597-025-06213-z.

Ran Zhao, Yanru Bai, Shuming Zhang, Jianing Zhu, Hongxing Liu, and Guangjian Ni. An open dataset of multidimensional signals based on different speech patterns in pragmatic Mandarin, 2025b. URL https://www.scidb.cn/en/detail?dataSetId=00fead6aa5dc4f2cb03d98fd29c5aad4. Dataset.

Bangyan Zhou, Xiaopei Wu, Zhao Lv, Lei Zhang, and Xiaojin Guo. A fully automated trial selection method for optimization of motor imagery based brain-computer interface. *PLOS ONE*, 11(9): e0162657, 2016. doi: 10.1371/journal.pone.0162657.

A Full proof of the proposition

Let subject s have EA whitener $W_s = \bar{R}_s^{-1/2}$, where $\bar{R}_s$ is the (arithmetic or Fréchet) mean reference covariance of domain s. EA maps a trial covariance C of domain s to $W_s C W_s^\top$.

MDM, single-domain (exact). Fix a held-out subject s and build class prototypes $P_c = \mathrm{mean}\{C_i : y_i = c,\ i \in s\}$ from s's own calibration trials. Under EA every such trial maps by W_s, so by congruence-equivariance of the Fréchet mean $P_c \mapsto W_s P_c W_s^\top$. A test trial X of subject s also maps by W_s. Congruence-invariance of AIRM gives $d(W_s P_c W_s^\top,\, W_s X W_s^\top) = d(P_c, X)$ for every c, so the full vector of class distances is unchanged and $\arg\min_c d(P_c, X)$ is identical with and without EA. Every prediction is preserved, hence $\Delta_{\mathrm{EA}} = 0$ exactly and 0 subjects change. □

MDM, cross-domain (no protection). If P_c is estimated from donor subjects $t \neq s$, then $P_c \mapsto W_t P_c W_t^\top$ while $X \mapsto W_s X W_s^\top$. The two arguments of $d(\cdot,\cdot)$ undergo *different* congruences, and $d(W_t P_c W_t^\top, W_s X W_s^\top) \neq d(P_c, X)$ in general; invariance provides no protection and the effect is empirically large.

Tangent-space logistic regression (approximate). Map each covariance to the tangent space at a reference M: $u(C) = \mathrm{vec}_{\sqrt{2}}\big(\log(M^{-1/2} C M^{-1/2})\big)$, where $\mathrm{vec}_{\sqrt{2}}$ stacks the diagonal and $\sqrt{2}$ times the strict upper triangle. The $\sqrt{2}$ is not cosmetic: it makes $\mathrm{vec}_{\sqrt{2}}$ a linear isometry from $\mathrm{Sym}(C)$ under the Frobenius inner product onto $\mathbb{R}^{C(C+1)/2}$, so that $\|u(C)\|_2 = \|\log(\cdot)\|_F$. Without it the induced coordinate map is a general similarity rather than an orthogonal one and the ℓ_2 argument below does not close. The released implementation uses the weighted map; the unweighted vech appears only in the deliberately affine-*sensitive* Euclidean rungs. Two facts follow. (i) If M is estimated in the *donor* domain but the test trial is in the target domain, EA does not cancel between reference and trial and a residual transform remains—a reference leak (prediction P1). (ii) Even with a shared in-domain reference, EA induces a fixed orthogonal rotation Q of the tangent coordinates, $u \mapsto Qu$. An L_2-regularised linear classifier is equivariant to Q (the optimum rotates with the features and the decision is preserved), so it remains exactly invariant; but a per-coordinate StandardScaler applies a diagonal rescaling D that does not commute with Q, breaking the invariance (prediction P2). Both predictions are confirmed in §6.2.

B Exactly which decoders the certificate covers

The proposition of §4 is a one-way implication, and its naive converse is false: a constant classifier has $\Delta_{\mathrm{EA}} = 0$ and is invariant to nothing. The direction that is valid is the contrapositive, and it is the one the diagnostic of §8 uses—if alignment changes a prediction, the pipeline is not (affine-invariant with an in-domain reference). A genuine biconditional is available once the quantifier is placed correctly.

Write $\mathcal{S}$ for a labelled reference set and X for a test covariance, all in $\mathrm{SPD}(C)$, and let $g(\mathcal{S}, X)$ be a decision rule. Let G act by congruence, $C \mapsto ACA^\top$ for $A \in GL(C)$, applied simultaneously to every element of $\mathcal{S}$ and to X (the single-domain case).

Proposition 2 (Characterisation)**.** *g satisfies $g(A \cdot \mathcal{S}, A \cdot X) = g(\mathcal{S}, X)$ for every $A \in GL(C)$, every $\mathcal{S}$ and every X* ***if and only if*** *g factors through the quotient by G—that is, there is a map $\tilde{g}$ on G-orbits with $g = \tilde{g} \circ \pi$, where π sends a configuration to its congruence orbit.*

Proof. ($\Leftarrow$) If $g = \tilde{g} \circ \pi$ then g is constant on orbits, and $(A \cdot \mathcal{S}, A \cdot X)$ lies in the same orbit as $(\mathcal{S}, X)$. ($\Rightarrow$) Suppose g is invariant. Define $\tilde{g}$ on orbits by $\tilde{g}(\pi(\mathcal{S}, X)) := g(\mathcal{S}, X)$. This is well defined exactly because any two representatives of an orbit differ by some $A \in GL(C)$, on which g agrees by hypothesis. Then $g = \tilde{g} \circ \pi$. □

What this does *not* give, and a claim we withdraw. An earlier draft asserted that the pairwise AIRM distances among $\mathcal{S} \cup \{X\}$ are a generating set of invariants for the congruence action, and concluded a clean "certified class". **That is false and we retract it.** At $C = 2$, the configurations $(I, \mathrm{diag}(e, e^{-1}))$ and $(I, \mathrm{diag}(e^{\sqrt{2}}, 1))$ have identical AIRM distance matrices ($d^2 = 2$) and different generalised spectra, hence lie in different orbits; $\det C_2 / \det C_1$ is a congruence invariant separating them and is not a function of the distance matrix. A dimension count makes the shortfall general: n points in $\mathrm{SPD}(C)$ have an orbit space of dimension $n\,C(C+1)/2 - C^2$ against only $n(n-1)/2$ pairwise distances. One consequence is concrete: MDM under the Stein/S-divergence is exactly congruence-invariant and is not a function of AIRM distance, so it belongs to the invariant class while falling outside any distance-based description of it.

A second gap is sharper still. Proposition 2 quantifies over *every* $A \in GL(C)$, whereas EA applies one *data-dependent* map $W = \bar{R}(\mathcal{S})^{-1/2}$. Because EA is idempotent—after alignment the reference is I, so $\mathrm{EA} \circ \mathrm{EA} = \mathrm{EA}$ (verified to 4×10^{-15})—the composite $h \circ \mathrm{EA}$ is exactly EA-invariant for *any* rule h whatsoever, including a Euclidean vech classifier with a per-feature scaler. "Exactly EA-invariant" is therefore strictly weaker than "$GL(C)$-invariant", and only the latter is characterised by Proposition 2. We state both gaps here rather than leave them, and we no longer claim a characterisation of the decoders the empirical certificate covers. What survives is Proposition 2 itself, the contrapositive diagnostic, and the empirical observation that a mean-free rule (k-NN) shows the same exact zero.

A caution. It is tempting to argue this through convexity, since congruences carry geodesically convex sets to geodesically convex sets on a Hadamard manifold. That is true and it is the wrong tool: preserving convexity is strictly weaker than being an isometry. Nor are the decision regions themselves convex—MDM's regions are Riemannian Voronoi cells, which are not geodesically convex in general. An isometry maps the Voronoi diagram to the diagram of the mapped prototypes, and that is all the proof requires.

C THE SAME STATEMENT IN VARIANCE UNITS

The proposition above is about a decoder. The same two facts give a statement about the *data*, with no classifier in it. For trial covariances C_i with subject $s(i)$, define the AIRM Fréchet variances

$$V_{\mathrm{total}} = \tfrac{1}{N}\sum_i d(C_i, M)^2, \quad V_{\mathrm{subj}} = \tfrac{1}{N}\sum_s n_s\, d(M_s, M)^2, \quad V_{\mathrm{within}} = \tfrac{1}{N}\sum_i d(C_i, M_{s(i)})^2,$$

with M and M_s the global and per-subject Fréchet means.

Proposition 3. *Under per-subject whitening by the* Fréchet *mean—that is, Riemannian Alignment,* ***not*** *the arithmetic-mean EA of He & Wu (2020)—$V_{\mathrm{subj}} \to 0$ exactly, and V_{within} is unchanged.*

The restriction is not cosmetic. The proof uses congruence-*equivariance* of the Fréchet mean, which the arithmetic mean does not have, so under arithmetic-mean EA the subject means do not all map to I and V_{subj} is *not* exactly zero. The measurements below are therefore of RA, and we label them as such.

Proof. EA maps subject s by $W_s = M_s^{-1/2}$. By congruence-equivariance of the Fréchet mean, $M_s \mapsto W_s M_s W_s^\top = I$ for every s, so all subject means coincide, the global mean is I, and $V_{\mathrm{subj}} = 0$. By congruence-invariance of d, $d(W_s C_i W_s^\top, W_s M_s W_s^\top) = d(C_i, M_s)$ termwise, so V_{within} is unchanged. □

Measured on all ten configurations, V_{subj} falls to between 10^{-7} and 10^{-19} of its original value—zero to the precision of the iterative Karcher mean—and V_{within} moves by at most 10^{-9} in absolute terms. Before alignment, V_{subj} is 29–68% of total variance.

Two cautions this table earns. First, the decomposition $V_{\text{total}} = V_{\text{subj}} + V_{\text{within}}$ is exact only in a flat space. On a Hadamard manifold the difference is non-negative *by theorem*—Sturm's NPC variance inequality applied per subject and summed gives $V_{\text{total}} \geq V_{\text{within}} + V_{\text{subj}}$ (Sturm, 2003)—so only its *magnitude* is an empirical question. We measure it at $+0.3\%$ of total at $C = 3$ rising to $+7.7\%$ at $C = 60$. An earlier draft presented the sign as something we had declined to assume; that was a gap in our reading, not caution. Second, the relevant sanity check on the decomposition is whether its class term tracks what a decoder can do: within-subject, the class share of within-subject variance correlates with within-subject balanced accuracy at Spearman $\rho = 0.90$ ($n = 10$ corpora, $p = 4 \times 10^{-4}$). We report that as a validation of the decomposition, not as a finding—both quantities measure class separability, so their agreement is in part definitional.

D Provenance of the Gowda montage subset

The source release provides a single 22-channel montage (Gowda et al., 2025; Gowda & Miller, 2024); the 6-channel variant is a channel subset selected by us, so the two Gowda rows are *not* independent corpora and any channel-count comparison across them is within-corpus. The subset is a fixed, contiguous block of six electrodes chosen under an external sensor-placement constraint, not to maximise accuracy: it is applied identically to every subject and trial before any analysis and never re-selected per subject or per fold, but the constraint was refined using per-channel signal-to-noise ratios and a 36-way decode comparison measured on three of these twelve subjects, so it is *not* independent of the reported data. The direction is conservative—the retained subset is the one those comparisons rejected, scoring 0.34 against 0.62 for the full montage—but the 6- and 22-channel rows differ in channel *quality* and in n/p regime as well as in count, and we draw no channel-count conclusion from their difference. The supplement records the index list, the selection evidence and the gaps in both. The Gowda source does not report a session count; we treat each subject as a single session.

E The covariance-distance cell

(3) Covariance distance was not shown to predict transfer failure—and we no longer call this a null. Three of six cells returned NaN and are excluded: a failed correlation is not a negative result. Of the three that computed, none reaches significance in the predicted direction, but the intervals make clear why: EMG-UKA $\rho = -0.40$, 95% CI $[-0.98, +0.91]$ on $n = 4$; Gowda $\rho = -0.30$, $[-0.75, +0.33]$ on $n = 12$; Shin $\rho = -0.10$, $[-0.47, +0.30]$ on $n = 26$. With intervals that wide the hypothesis is *untested*. We also note against ourselves that Shin's *partial* Mantel is significant ($r = +0.27$, $p = 0.036$), which is the statistic the claim is really about and which points the other way. We report the cell and draw no conclusion from it.

F The displacement intervention

We apply a synthetic channel-mixing of magnitude δ to the held-out subject's EMG and toggle EA on/off (Figure 3). Without alignment, accuracy collapses toward chance; with alignment it holds, up to a breakdown point where the mixing becomes ill-conditioned (max gain $+0.351$ on GRABMyo at $\delta = 0.9$, with recovery failing beyond $\delta = 0.7$; replicated on Gowda, max $+0.222$). The curve is an inverted U: near zero where transfer is already good, maximal at intermediate displacement, collapsing when the displacement is too large for any congruence to undo. This is an intervention, not a correlation—direct causal evidence that EA removes the electrode-displacement nuisance it is claimed to. We report the recovery curve rather than a single effect size, and flag this as the one result here without a per-subject confidence interval.

G Recolouring versus per-channel standardisation

An earlier version of this work reported that EA, Riemannian Alignment and CORAL are "interchangeable". Re-run at a declared regularisation ($\alpha = 0$, §6.4) and scored with two one-sided tests against a pre-registered margin of 0.02, that claim does not survive, and we retract it. What survives is the control it was meant to support.

Surviving. EA $-$ z-score $= +0.033$, equivalence rejected on all three corpora where both arms are defined, and under the log-map at $\alpha = 0$ per-channel z-scoring is *actively harmful* (-0.015).

Covariance recolouring genuinely beats per-channel standardisation, in both featurisers, at every α.

Not established. EA $\approx$ CORAL is equivalent on $1/3$ corpora and inconclusive elsewhere; EA $\approx$ RA on $\mathbf{0/5}$. The verdict is also featuriser-conditional—under the log-map, RA *beats* EA on both GRABMyo and Shin—so the choice between the Euclidean and Fréchet mean is **unresolved**, not irrelevant. Table 11 in Appendix O gives the contrasts.

H THE REFERENCE-SIZE REGIME IN FULL

Alignment is not free at small reference sizes (Figure 4). We determine the harmful regime from *intervals* on the per-subject gain at each reference size, not from the sign of a mean. On GRABMyo ($n = 43$, baseline 0.761 against a chance of 0.167), the gain interval lies strictly *below* zero at 2, 4 and 8 reference trials and first clears zero at 16: below that point the whitener injects more estimation noise than it removes nuisance. This is a practical result—a reference estimated from a handful of calibration trials is worse than no alignment at all—and it is the regime in which per-domain whitening acts as a conditioner rather than a domain correction.

It replicates, on a minority. Across all nine corpora the harmful regime appears on **three**: GRABMyo (harmful at 2, 4, 8; clearing at 16), BCI IV-2b and Weibo2014 (harmful at 2, clearing at 4). Three show none; CSL's baseline is at chance and cannot speak to it; two are equivocal (BCI IV-2a is harmful at 4 while clearing zero at 2—a non-monotone curve, not a threshold—and the two Gowda montages disagree). GRABMyo's is much the most extensive. The phenomenon is real, not universal, and we report the minority rather than the headline.

Does the knee scale with channel count? We cannot say, and we withdraw our earlier claim that it does not. Across five channel counts the correlation between knee and C is $r = -0.000$, but the Fisher-z 95% interval on $n = 5$ points is $[-0.88, +0.88]$—an interval containing nearly every correlation there is. A null estimate at that precision is *untested*, not refuted, and reporting it as a refuted law was an error of exactly the kind this paper is about. Two of the five points are moreover the same corpus at two montages (Table $1^{\dagger}$), so the effective n is at most four.

I ALIGNMENT BY MODALITY, IN FULL

A widely-cited intuition holds that alignment helps EMG but not EEG. On the only well-powered multi-channel EEG set (Shin, $C = 30$), EA *helps* by $+0.101$ (95% CI $[+0.065, +0.132]$, $p = 1.6 \times 10^{-5}$); the EMG sets improve comparably (Table 6).

Across the declared family of $m = 10$ corpora, GRABMyo, Shin and Gowda-22 survive Holm–Bonferroni ($p_{\mathrm{adj}} = 9.4 \times 10^{-6}$, 1.4×10^{-4}, 2.7×10^{-2}). We give both statistics, because the corrected count understates the consistency: **all ten point estimates are positive, eight of ten** 95% **intervals exclude zero, three survive correction over ten.** The gap is power—six corpora have $n \leq 12$, and at $n = 4$ the smallest attainable signed-rank p is 0.125, so two members cannot reach significance at any effect size. We keep them: dropping a member that cannot pass is how a family gets curated.

What the EEG baseline is, and why it strengthens the claim. Shin's cross-subject decoder clears chance but recovers only 12% of the margin its own within-subject decoder reaches (§3, Table 10): 0.397 against a chance of 0.333, where within subject it attains 0.858. The alignment gain reported above is therefore **larger than the entire unaligned margin over chance** —$+0.101$ against $+0.064$—and the same holds on EMG-UKA ($+0.077$ against $+0.046$). We state this rather than leave it to be computed, and we read it as the strongest available form of the paper's thesis rather than as a caveat: where the unaligned cross-subject decoder recovers almost nothing, the alignment step supplies more of the usable signal than the decoder does, which is difficult to reconcile with alignment as a mere preprocessing convenience. It also fixes the scale of the comparison in Table 6: "EA helps EEG as much as EMG" is a statement about gains measured against baselines that differ by a factor of six in how much they transfer (12% on Shin against 77% on GRABMyo), and the two gains are not interchangeable quantities merely because they are similar numbers.

Table 6: Alignment gain by corpus, with the Holm–Bonferroni correction over the declared family of $m = 10$ applied in the table rather than in the text. 'improved': subjects whose balanced accuracy rose.

dataset	modality	n	base	Δ_{EA}	95% CI	p	p_{adj}	improved
grabmyo*	forearm sEMG	43	0.761	$+0.071$	$[+0.047, +0.103]$	$9.4\mathrm{e}{-7}$	$9.4\mathrm{e}{-6}$	32/43
shin*	EEG	26	0.397	$+0.101$	$[+0.065, +0.132]$	$1.6\mathrm{e}{-5}$	$1.4\mathrm{e}{-4}$	23/26
gowda22*	orofacial sEMG	12	0.617	$+0.167$	$[+0.083, +0.231]$	0.003	0.027	11/12
weibo2014	EEG (motor imagery)	10	0.586	$+0.057$	$[+0.025, +0.092]$	0.020	0.137	9/10
gowda3	orofacial sEMG	12	0.569	$+0.086$	$[-0.008, +0.164]$	0.064	0.384	8/12
bnci2014004	EEG (motor imagery)	9	0.643	$+0.043$	$[+0.013, +0.093]$	0.074	0.384	7/9
bnci2014001	EEG (motor imagery)	9	0.685	$+0.041$	$[+0.008, +0.084]$	0.098	0.391	7/9
emguka	silent-speech sEMG	4	0.379	$+0.077$	$[+0.063, +0.086]$	0.125	0.391	4/4
zhou2016	EEG (motor imagery)	4	0.669	$+0.022$	$[+0.009, +0.037]$	0.125	0.391	4/4
csl	HD-sEMG	8	0.508	$+0.013$	$[-0.018, +0.030]$	0.250	0.391	6/8

*survives Holm–Bonferroni at FWER 0.05 over the declared family; the family is fixed in source before the p-values are read, and the members that fail are listed here rather than omitted.

Recolouring versus per-channel standardisation. Per-channel z-scoring is actively harmful under the log-map, and EA and Riemannian Alignment are *not* interchangeable—an equivalence we previously asserted and now withdraw, established on $\mathbf{0/5}$ corpora (Appendix G).

How much of the gain is the featuriser? A featuriser ablation attributes the aligned pipeline's advantage between the covariance representation and the alignment step itself; the split is reported in Appendix K.

J THE SESSION-VERSUS-USER CONTRAST IN FULL

The framing that motivates alignment is inter-*subject* variability, which predicts a larger benefit across users than across sessions. On GRABMyo the opposite holds: cross-user $\Delta_{\mathrm{EA}} = +0.070$, cross-session $+0.115$, and paired over the same 43 subjects the difference is $+\mathbf{0.045}$, 95% CI $[+0.012, +0.076]$, Wilcoxon $p = 0.003$.

It does not replicate. Four further corpora carry both splits and none reproduces the ordering: Shin $+0.014$ $[-0.021, +0.049]$, BCI IV-2a -0.000 $[-0.043, +0.026]$, BCI IV-2b -0.022 $[-0.071, +0.019]$, Zhou2016 -0.012 $[-0.044, +0.018]$. One of five, and the four failures do not point the same way as the one success.

The declared family, and the correction. The primary contrast is session − user at full n, one per corpus, fixed before the p-values were read; Holm over those $m = 5$ leaves GRABMyo surviving at $p_{\mathrm{adj}} = 0.017$ and the rest at 1.000. We also report two controls per corpus—the same contrast at matched training size, and the size effect on the user arm—and correct nothing over them, because they are not independent of the primary: **primary = matched + size** identically, and we verify the identity to four decimals on all five corpora. A reader who instead corrects over all fifteen printed intervals finds that none survives; we state that too, because reporting fifteen and correcting none is not an option we are willing to take.

We therefore make the weaker and better-supported claim: *the inter-subject motivation does not predict where alignment helps.* Not that alignment reliably helps more across sessions—that whether it does is not determined by the kind of domain boundary.

Two candidate confounds. **Headroom does not explain it**—the cross-session decoder starts *higher* (0.815 vs. 0.761) and still recovers more. **Estimation regime is not separated from domain type**: the arms train on 10,584 vs. 168 trials, i.e. $n/p = 77.8$ against 1.2 on the $p = 136$ `vech` featuriser, the near-singular regime of §6.7. Matching the sizes moves cross-user Δ_{EA} to $+0.090$, but neither component excludes zero (size $+0.020$ $[-0.009, +0.050]$; residual $+0.025$ $[-0.008, +0.060]$). *The total effect is established on one corpus; its decomposition is not.*

K THE FEATURISER ABLATION

Holding decoder and folds fixed and changing only the feature map, the EA gain on GRABMyo falls from $+0.071$ (`vech`) to $+0.028$ (log-map) at $\alpha = 0$—60% of what alignment appears to buy is the featuriser—while on Shin the same swap leaves 93% intact. The corpora disagree by 53 points, so we quote no pooled figure. The log-map at a *global* reference is not invariant to a per-subject congruence: it linearises the nuisance into something a linear boundary can partly absorb, and does not remove it.

L DESIGN IMPLICATIONS IN FULL

The diagnostic. If alignment changes anything in a pipeline you believe is affine-invariant, the pipeline is not invariant—and the leak is usually a reference borrowed from the wrong domain, or a per-feature scaler. Match the alignment to the model's invariance class rather than benchmarking variants against each other. We no longer claim the variants are interchangeable (Appendix G); the axes that demonstrably move the result are *where* the reference is estimated, *how many* trials estimate it, *which* featuriser follows, and *what* regularisation is applied. Those four change results by up to 0.26 and can flip the sign. Which mean builds the reference has not been shown to matter, and has not been shown not to.

M THREATS TO VALIDITY IN FULL

The classical TS+LR pipeline reaches 0.922 balanced accuracy cross-user on 43 subjects—our claims are about alignment, not deep architectures. Gowda has no reported session structure, so it cannot contribute cross-session evidence, and its 6- and 22-channel rows are the same 12 subjects under a montage subset we selected, so they are not independent observations of channel count. EMG-UKA has $n = 4$, which is below the size at which a one-sided signed-rank test can reach $p < 0.05$ at all; we report its cells and exclude them from every conclusion. Equivalence at our pre-registered margin of 0.02 is unreachable on four of five corpora at these sample sizes; we report the interval and decline to claim a null rather than reading an underpowered test as evidence of no difference.

N NEGATIVE RESULTS IN FULL

We report these because they cost real compute and correct the record—two refute hypotheses previously advanced in this line of work. **(1) No modality dissociation** (§6.3). The earlier claim rested on an $n = 3$ dataset whose result file was degenerate, and on a peri-auricular EEG extraction that had silently collapsed to a *single channel*, where EA is scalar rescaling. **(2) Affine invariance does not by itself explain the effect.** EA gives MDM—fully AIRM—$+0.105$ zero-shot; §4 gives the correct account (domain-mismatched reference, not classifier family). **(3) Covariance distance was not shown to predict transfer failure.** Three of six cells returned NaN and the three that computed have intervals spanning nearly every correlation there is; the hypothesis is *untested*, not refuted, and we report the cell in Appendix E rather than drawing a conclusion from it here.

An exclusion that is ours, not the data's. A sixth public corpus, T-MSPD (Zhao et al., 2025a;b), is at chance for every decoder we tried, and at chance *within subject* (0.101 against a chance of 0.100) where GRABMyo reaches 0.938 through identical code. Our probe localises the cause to the epochs rather than the decoder: no class information anywhere in the window (class F-ratio 1.05 against a shuffled null of 1.51) and envelope energy piling at both edges with a dead centre ($2.09\times$ uniform)—windows sitting *between* utterances. The source release provides per-trial event markers and hand-annotated speech segments, and our ingest used neither. We therefore report this as a failure of *our* pipeline and exclude the corpus pending a corrected ingest; it would be precisely the error this paper is about to read an at-chance number produced by our own windowing as a fact about the dataset.

Table 7: **Users served, against budget.** Fraction of subject-cells for which a total budget of T labelled target trials reaches the aligned cross-subject level, over all within-corpus cells. A cell fraction is a mean over subjects and a mean is not a user; the right-hand column is the share served by *no* tested budget.

corpus	T=10	T=20	T=30	T=40	T=60	T=80	T=120	never served
GRABMyo (sEMG, n=43)	59%	76%	80%	84%	87%	89%	90%	10%
BCI IV-2a (EEG-MI, n=9)	35%	54%	60%	65%	70%	74%	77%	22%

Table 8: MDM k-shot Δ_{EA} is exactly 0.000 on every subject of every configuration ($\mathbf{0/137}$ changed). Broken-invariance rungs are all large; the ordering *within* the broken rungs is not monotone, so we claim the dichotomy and the approximate middle rung, not a total order.

dataset	C	n	MDM k-shot	TS+LR k-shot	TS+LR 0-shot	MDM 0-shot	Euclid
bnci2014004	3	9	$+\mathbf{0.000}$	$+0.006$	$+0.032$	$+0.023$	$+0.044$
emguka	6	4	$+\mathbf{0.000}$	-0.003	$+0.100$	$+0.078$	$+0.077$
gowda3	6	12	$+\mathbf{0.000}$	$+0.017$	$+0.089$	$+0.117$	$+0.072$
zhou2016	14	4	$+\mathbf{0.000}$	$+0.001$	$+0.047$	$+0.037$	$+0.020$
grabmyo	16	43	$+\mathbf{0.000}$	-0.003	$+0.030$	$+0.085$	$+0.070$
bnci2014001	22	9	$+\mathbf{0.000}$	-0.002	$+0.047$	$+0.121$	$+0.041$
gowda22	22	12	$+\mathbf{0.000}$	-0.006	$+0.156$	$+0.150$	$+0.178$
shin	30	26	$+\mathbf{0.000}$	$+0.008$	$+0.095$	$+0.038$	$+0.100$
csl	35	8	$+\mathbf{0.000}$	-0.008	$+0.011$	-0.024	$+0.010$
weibo2014	60	10	$+\mathbf{0.000}$	$+0.016$	$+0.089$	$+0.109$	$+0.060$

O ADDITIONAL TABLES AND FIGURES

The informativeness floor. A zero is evidence only where the classifier classifies. Declaring the threshold in advance—informative if raw accuracy clears chance by 0.05, marginal above 0.02—the rung is at chance on Weibo2014 (0.515 vs 0.500, $C = 60$) and marginal on CSL (0.542, $C = 35$): 10 fits at chance, 8 marginal, so the certificate rests on $\mathbf{119/137}$ informative fits and we quote both. The two weak cells are the two highest-dimensional ones, which is this paper's own mechanism in its own measurement—the rung builds each prototype from `cal_k` $= 5$ covariances, and at $C = 60$ that is five points locating a mean on a 1830-dimensional manifold.

Table 9: Per-corpus certificate. `changed` is the number of subject-fits whose MDM k-shot prediction set was altered by EA. † marks the two configurations whose k-shot rung sits within 0.05 of chance: their zeros are algebraically valid and empirically uninformative, and they are excluded from the 119.

corpus	C	K	chance	MDM k-shot	fits	changed
BCI IV-2a	22	2	0.500	0.615	9	0
BCI IV-2b	3	2	0.500	0.559	9	0
CSL †	35	2	0.500	0.542	8	0
EMG-UKA	6	3	0.333	0.572	4	0
Gowda-22	22	3	0.333	0.867	12	0
Gowda-6	6	3	0.333	0.756	12	0
GRABMyo	16	6	0.167	0.771	43	0
Shin	30	3	0.333	0.679	26	0
Weibo2014 †	60	2	0.500	0.515	10	0
Zhou2016	14	2	0.500	0.626	4	0
total					**137**	**0**

Table 10: Cross-subject transfer, and what each alignment gain is measured against. *within*: within-subject balanced accuracy. *LOSubO*: leave-one-subject-out. *kept*: the fraction of that corpus's own within-subject margin over chance that survives crossing subjects. A corpus can clear chance cross-subject and still keep almost none of what is there; the two right-hand columns are not interchangeable and we report both.

corpus	C	within	LOSubO	chance	kept	Δ_{EA}
emguka	6	0.746	0.379	0.333	11%	+0.077
shin	30	0.858	0.397	0.333	12%	+0.101
csl	35	0.557	0.513	0.500	22%	+0.013
gowda22	22	0.889	0.617	0.333	51%	+0.167
weibo2014	60	0.656	0.586	0.500	55%	+0.057
gowda3	6	0.722	0.569	0.333	61%	+0.086
zhou2016	14	0.777	0.670	0.500	61%	+0.022
bnci2014004	3	0.690	0.643	0.500	75%	+0.043
grabmyo	16	0.939	0.762	0.167	77%	+0.071
bnci2014001	22	0.728	0.685	0.500	81%	+0.041

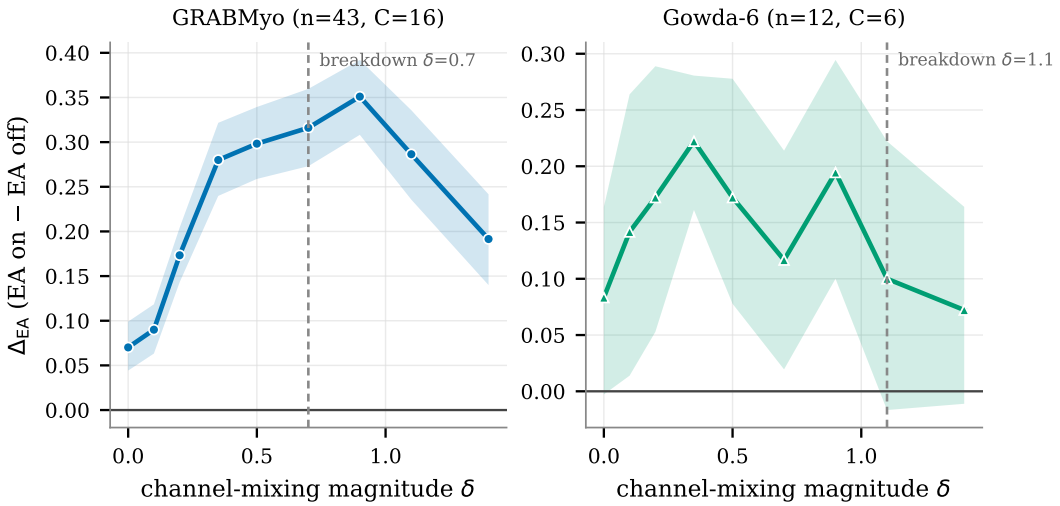

Figure 3: Alignment rescues electrode displacement. Balanced accuracy vs. displacement magnitude δ, alignment off (grey) vs. on (teal); the shaded band is the EA gain. Left: GRABMyo ($n = 43$); right: Gowda ($n = 12$).

P Riemannian Procrustes Analysis on the budget axis

Riemannian Procrustes Analysis (Rodrigues et al., 2019) is often described as a stronger alignment than EA, and therefore as the natural baseline against which a paper about alignment should be read. The description elides a distinction that matters here. RPA is a composition of three operations, and they do not carry the same supervision cost:

- **recentering** translates each domain's centre of mass to the identity. Unsupervised. Under the Fréchet mean it is exactly Riemannian Alignment, which we measure here as its own zero-label arm—the RA column of Table 12—so it is priced in this appendix rather than assumed away. We do *not* lean on any equivalence between EA and RA: Appendix G finds that claim established on $\mathbf{0/5}$ corpora and retracts it, with the sign flipping under the log-map. The statement here is therefore "measured, and cheap on these corpora," not "known to be free."

Table 11: Alignment-family contrasts at $\alpha = 0$, `vech` featuriser, TOST against a pre-registered margin of 0.02. "Established on k/n" counts corpora whose 90% interval lies inside $(-0.02, +0.02)$. Inconclusive is not refutation.

contrast	mean	equivalence
EA − z-score	$+0.033$	rejected on $3/3$ — **the surviving control**
EA − CORAL	-0.008	established on $1/3$ (GRABMyo); inconclusive elsewhere
EA − RA	-0.008	established on $\mathbf{0}/\mathbf{5}$; sign flips under the log-map

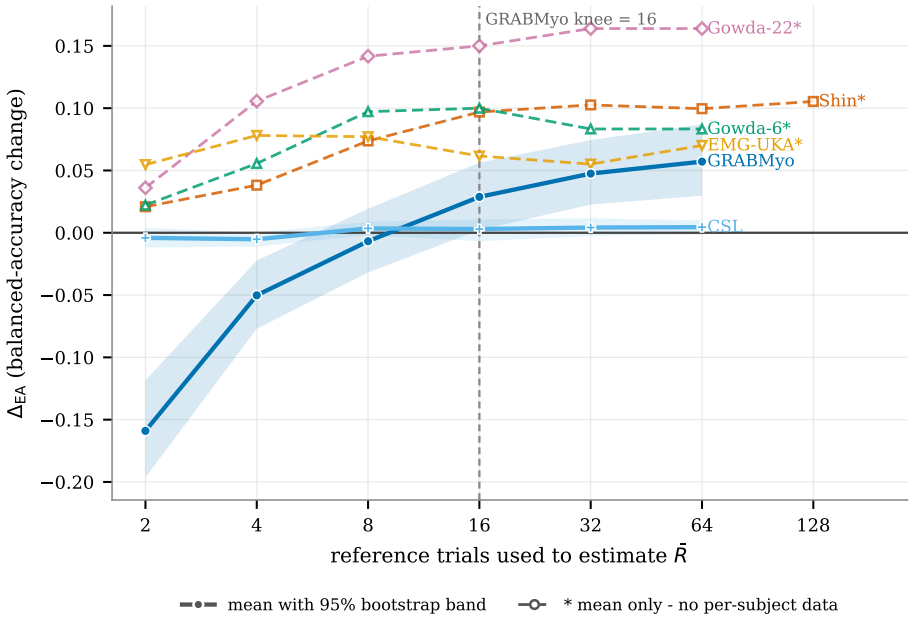

Figure 4: Reference-size dose–response: Δ_{EA} vs. number of reference trials (log scale). GRABMyo (black) has the most extensive harmful small-reference regime: its gain interval lies below zero at 2, 4 and 8 reference trials and first clears zero at 16. CSL is a different case and is not among the three—its gain never clears zero at any reference size, so it has no small-reference *regime*; it simply does not benefit from alignment. The rest are positive at every size tested. Thresholds are read from intervals, not from the point estimates plotted here.

- **rescaling** equalises each domain's dispersion about its mean. Unsupervised, and measured here for the first time in this work; the gap between the RA and $\mathrm{RPA}_{\mathrm{unsup}}$ columns of Table 12 is what it buys.
- **rotation** matches class means *between* the source and target domains by optimisation on the Grassmann manifold. This step reads target labels. In the reference implementation the fit signature is `TLRotate.fit(X, y_enc)` and the documented behaviour is to rotate each source domain "so to match its class means with those from the target domain".

The consequence is structural, not incidental. RPA-full is *not* a zero-shot alternative to EA that belongs beside it in the aligned column: two of its three steps cost no labels at all and are reported here as zero-label arms, while the third spends the very resource this paper denominates in. **RPA-full is a competitor on the budget axis of §6.5, and the only coherent way to benchmark it is to price it.** That is what this appendix does.

Design, and why the label accounting is the whole experiment. Unlabelled target data is free—the recording already exists—whereas labels are the scarce resource. Accordingly, at every budget k the recentering and rescaling are fit on the *full unlabelled* target calibration pool, and *only* the rotation sees the k labelled trials per class. Every spent label is thus attributable to the rotation, and

$k = 0$ is a genuine zero-shot arm (recentre + rescale, no labels). This is not a stylistic choice: fitting the recentering on k trials instead leaves the rotation ill-conditioned, because a rotation that matches class means is only meaningful once both domains have been brought to a common centre.

A convergence audit, because the result depends on it. The Grassmann optimiser emits a non-convergence warning on essentially every domain fit (4296 across the reported runs). We report this rather than suppress it, and we tested whether it explains RPA's showing, because if it did then this benchmark would be an artifact rather than a measurement. It does not. Loosening the step tolerance by five orders of magnitude changes neither the accuracy nor the runtime, and *sextupling* the iteration budget—a six-fold increase in wall time—leaves the resulting balanced accuracy unchanged to three decimal places on two of the three subjects audited and very slightly *lower* on the third. In no case does additional optimisation improve RPA's showing, which is the objection the audit was run to answer. The optimiser has reached a plateau on which the descent-direction norm never falls below its threshold while the solution itself no longer moves; the warning is cosmetic. The audit script and its log are released with the code. We note this at length because an earlier draft of this appendix asserted that a design change had eliminated the non-convergence: that claim was false, produced by a diagnostic that captured warnings in the parent process while the optimiser ran in worker subprocesses, and it is withdrawn here.

Every arm is recomputed in a single harness on *identical* fixed test blocks, held out per subject before any k is chosen. We do not quote a level measured by another script: the whole point of §6.5 is that a difference and a level are not interchangeable, and levels produced by two harnesses are not either. We are also deliberately stricter than the field on one axis—every target-side reference (EA, RA, recentering, rescaling) is estimated on the calibration pool only, never on the test block, whereas the conventional EA implementation is transductive and uses all target trials. Levels here are therefore not comparable to the published EA numbers or to our own Table 3; only the within-table contrasts are.

Settings, and an asymmetry we are not hiding. We use the reference implementation in `pyriemann` 0.12 (`TLCenter`, `TLScale`, `TLRotate`) rather than a reimplementation, because an unfaithful substitute for the Grassmann optimisation—an SVD–Procrustes stand-in, say—would produce a strawman rather than a benchmark. Rotation uses the package defaults (`metric=euclid` for matching class means, `maxiter`$= 10^4$, `tol_step`$= 10^{-9}$), and 6 calibration draws are averaged per cell.

The comparison in the lower panel is *not* like-for-like in data quantity, and should not be read as one. At budget k, RPA-full sees k labelled target trials *plus* the unlabelled target pool *plus* the donor corpus, whereas the in-domain arm sees only the k labelled trials. The two are matched on the resource we claim is scarce—labels—and on nothing else. This is the deployment question (collect k trials, or borrow donors and spend k trials on a rotation?), not a controlled comparison of estimators, and a semi-supervised in-domain method that also exploited the unlabelled pool would be a third arm we do not run.

Three findings, ordered by how well they replicate. First, and robustly across both corpora and both decoders: **the rotation must be paid for long before it pays, and on this evidence it is a bad deal.** RPA-full does not exceed RPA's own *unsupervised* $k = 0$ level until roughly 40–60 labelled trials per class—80 to 120 total labels—and in 1 of the cells it never does so within the tested range at all. Below that point the rotation is *actively harmful*: it degrades a pipeline that recentering and rescaling had already improved for nothing. This is the same shape as the reference-size regime of §6.7 arising from a different mechanism—a rotation fit to class means estimated from a handful of target trials is fit to noise, and it destroys the donor structure that recentering had made usable. It is a second instance of the paper's general point: an alignment operation has an operating regime, and outside it the operation is not merely useless but negative. Note that this comparison is *conservative* in RPA's favour, because the RPA-full arm's classifier also trains on the k labelled target trials while the $k = 0$ arm does not; the rotation is harmful enough to overwhelm that advantage.

Second, on the deployment question the answer is closer than either the reviewer's framing or ours would suggest. RPA-full does exceed in-domain k-shot at matched budget, in 10 of 10 corpus $\times$ k cells under MDM and 7 of 10 under TS+LR, but by margins that are mostly small, and the advantage is concentrated exactly where in-domain data is weakest: it is largest on BCI IV-2b, the corpus with

Table 12: **Riemannian Procrustes Analysis priced on the budget axis.** RPA's recentering and rescaling are unsupervised; its rotation matches class means *between* domains and therefore consumes labelled target trials. Upper panel: the zero-label arms. Lower panel: the rotation's price, and RPA-full against in-domain k-shot at matched budget. All arms recomputed on identical fixed test blocks; every target-side reference is fit on the calibration pool only, never on the test block. Point estimates over subjects (SD across subjects in the text); no bootstrap over subjects, unlike the main-text tables.

corpus	decoder	no align.	EA	RA	$\text{RPA}_{\text{unsup}}$	RPA−EA
BCI IV-2a	MDM	0.620	0.732	0.735	0.747	+0.015
BCI IV-2a	TS+LR	0.688	0.731	0.733	0.738	+0.006
BCI IV-2b	MDM	0.663	0.689	0.692	0.692	+0.003
BCI IV-2b	TS+LR	0.649	0.683	0.683	0.682	-0.001

budget axis: RPA-full / in-domain k-shot, balanced accuracy

corpus	decoder	k=5	k=10	k=20	k=40	k=60
BCI IV-2a	MDM	0.649/0.638	0.667/0.657	0.695/0.694	0.721/0.714	0.752/0.746
BCI IV-2a	TS+LR	0.634/0.631	0.659/0.661	0.689/0.700	0.715/0.731	0.738/0.734
BCI IV-2b	MDM	0.652/0.628	0.664/0.635	0.678/0.643	0.688/0.667	0.690/0.666
BCI IV-2b	TS+LR	0.651/0.623	0.661/0.638	0.672/0.657	0.687/0.680	0.688/0.685

the largest $k^\star$ (> 240 total labels), where donor pooling yields a better-conditioned prototype than the single-subject data buys back inside our grid—not than any amount of it, since 4 of its 9 subjects do cross their own aligned level by $k = 80$ (Table 3). Donor data plus a supervised rotation is therefore a genuine competitor to spending the same labels in-domain, and which wins is corpus-dependent. That is a result about *how* to spend a budget, and it presupposes rather than undermines the claim that the budget is the right axis.

Third, and *not* replicated: the zero-shot column may not be saturated by EA. RPA's unsupervised half—recentre + rescale, no labels at all—exceeds EA on BCI IV-2a under both decoders with bootstrap intervals excluding zero (+0.015, [+0.006, +0.025] for MDM; +0.006, [+0.001, +0.010] for TS+LR; 6/9 and 6/9 subjects respectively; exact two-sided Wilcoxon $p = 0.031$ and $p = 0.047$ at $n = 9$, *uncorrected* across the four corpus × decoder cells tested—under the Holm–Bonferroni rule §3 applies to our primary family, neither survives, since the smaller p already exceeds $0.05/4$). On BCI IV-2b it does not: the interval spans zero under MDM (+0.003, [-0.002, +0.011]) and the point estimate is negative under TS+LR (-0.001, [-0.005, +0.003]). One corpus of two is a hypothesis, not a finding, and we report it as such. Were it to hold more widely it would *raise* the price of the aligned column rather than lower it—the level a practitioner must buy with labels would be higher than Table 3 assumes, making the $k^\star$ reported there conservative—but two corpora cannot establish that, and we do not claim it.

What this is not. Two corpora, both binary motor imagery, 18 subject-cells, 6 calibration draws per cell, point estimates over subjects for the k-curves with no bootstrap—weaker evidence than the main-text tables, and reported as such. Within-cell variation across calibration draws averages 0.030 balanced accuracy, which is why single-draw numbers are not reported at all. The two nominally-significant zero-label effects, +0.015 and +0.006, both lie *inside* the 0.02 practical-equivalence margin this paper pre-registers elsewhere, and are of the same order as that across-draw SD. We therefore assert nothing about the world from them, and report them only because a zero-label arm that beats EA at all is the kind of result a reader should be able to see and re-test.

Zhou2016 was attempted and is excluded, for two independent reasons. It is already excluded from pricing in the main text (§6.5) because its alignment-gain interval spans zero, so there is no gain on that corpus for a competing method to be priced against. It is also numerically unsuitable here: 2 of its 1199 covariance matrices have a smallest eigenvalue of -3.4×10^{-15} and the worst condition number exceeds 10^{302}, so the Fréchet mean underlying every recentering step is not defined without a regularisation constant—and §6.4 is a section about a regularisation constant changing a conclusion, which is not a knob we will turn silently to add a corpus.

The EMG corpora are absent, and their absence is not arbitrary: their $k^\star$ sits at the grid floor, so no competing method's curve is resolvable there at all, whereas the corpora used here carry the only priced gains large enough for a curve to be visible. TSMNet (Kobler et al., 2022) remains unbenchmarked, for a reason worth stating: it is an end-to-end architecture rather than an alignment operator, so adding it as a row would vary the decoder and the alignment simultaneously—precisely the confound the factorial of §6.2 exists to break. The right TSMNet experiment is an additional factor level, not an additional row, and we do not have it.

Q WHAT $k^\star$ IS, AND HOW WIDE IT IS

$k^\star$ as tabulated in Table 3 is the first grid point at which the *mean* in-domain curve reaches the *mean* aligned level. It is therefore a crossing of two averages: not an interval, and not any individual subject's budget. The paper's headline quantity is denominated in it, so we report here what it conceals rather than let the integers stand.

Subject bootstrap. Resampling subjects with replacement (4000 draws) and re-reading the crossing inside each draw, as recorded by the run that produced Table 3: on BCI IV-2a the point estimate $k^\star = 60$ carries a 95% interval of $[20, 60]$, with 22% of draws not crossing anywhere in the grid; on BCI IV-2b the interval is $[40, 120]$ with 77% of draws not crossing; on Zhou2016, $[10, 60]$ with 57% not crossing. An interval computed only over the draws that cross is conditional on crossing, so the never-crossing fraction is reported beside it in every case and neither number should be read without the other. Where that fraction exceeds roughly a fifth, $k^\star$ is a lower bound and nothing more.

Per-subject crossings, which are more informative than the mean. Reading each subject's own in-domain curve against *that subject's* aligned level, from the per-subject records released with this paper:

- **BCI IV-2a**: crossings at $k = 5, 10, 20, 20, 40$; 4 of 9 subjects never cross. The median crosser is at 20 and the latest is at 40—so the tabulated $k^\star = 60$ is *later than every subject who crosses at all*. The mean curve crosses late because the four non-crossers hold it down, not because any subject needs 60 trials per class.
- **BCI IV-2b**: crossings at $k = 20, 40, 60, 80$; 5 of 9 never cross. Four subjects therefore do reach their own aligned level within the grid, which is why the $k^\star$ for this corpus is reported as censored at > 120 per class and not as unbounded.
- **Zhou2016**: crossings at $k = 10$ and $k = 60$; 2 of 4 never cross.

What follows. Every $k^\star$ in this paper should be read as a point estimate with roughly a factor-of-three interval and a censored tail. The direction of the claims built on it is unchanged—five labelled trials per class do not recover the aligned level on any binary corpus, and the budget is tens of trials rather than a handful—but the precision a bare integer suggests is not there, and a mean crossing can sit outside the range of every individual it averages. That is not a defect of the measurement so much as the reason this paper reports a per-user service curve (Table 7) alongside the corpus-level number, and declines to offer a single constant. A practitioner provisioning for a population needs the fraction served at a budget, which the service curve gives; a practitioner provisioning for one user needs that user's own crossing, which only calibration on that user can give. The corpus mean answers neither question exactly, and we no longer present it as though it answered both.