NO SINGLE RECIPE FOR EEG: PHASE, TASK REGIME, AND DATA BUDGET INTERACT

Anonymous authors
Paper under double-blind review

ABSTRACT

For designers of electroencephalography (EEG) foundation models, input choice depends on the target task and available labeled-data budget. We compare raw patches, Morlet log power, and complex coefficients on P300 and N170 event-related potential tasks and one-second sleep-segment classification across five labeled-subject counts and three seeds. We hold the transformer body, subject splits, epochs, and optimization schedule fixed; only the front end varies. Token counts and compute costs differ.

On the ERP composite, raw patches outperform fixed Morlet log power throughout the tested range, with a 0.09–0.12 balanced-accuracy advantage after matching normalization scope. Matching token count and per-example compute in both directions retains the log-power advantage at 20 and 40 subjects; the advantage diminishes at larger counts. Complex coefficients from the same filterbank reduce the evoked deficit. In the ERP composite, the largest decomposition contrast is per-trial, per-frequency phase rotation: removing it improves balanced accuracy by 0.12–0.20. Frame averaging changes ERP-composite accuracy by at most about 0.006 at the same output rate. The rotation preserves pre-normalization magnitudes but changes event alignment and cross-frequency relationships.

A tested convolutional spectral embedding does not remove the evoked deficit. With native-target masked pretraining, raw gains more over scratch than complex inputs, although the arms predict different targets. Under a shared target, the six-seed experiment establishes neither a directional advantage nor equivalence between their gains.

1 INTRODUCTION

An EEG foundation-model designer must choose which signal features the model receives. Raw samples retain the waveform; log power summarizes local spectral energy but discards phase; complex coefficients retain phase as real and imaginary components. Classical EEG pipelines use spectral energy, as do influential convolutional networks (Schirrmeister et al., 2017; Lawhern et al., 2016). Pretrained EEG models use raw inputs (REVE and EEGPT (Ouahidi et al., 2025; Wang et al., 2024a)), spectral inputs (BIOT (Yang et al., 2023)), or combined temporal-spectral inputs (CBraMod (Wang et al., 2024b)); each commits to one front end.

For a designer building for a particular EEG task, the key question is which input performs better with the labeled subjects available. Rankings across complete models cannot isolate this choice: architecture, pretraining target, corpus, and input often change together. Controlled comparisons offer a closer test. CBraMod ablates its temporal and frequency branches, and TFM-Tokenizer compares raw, short-time Fourier transform (STFT), and combined inputs (Wang et al., 2024b; Pradeepkumar et al., 2026). These studies still leave open how front-end differences change with subject count and which operations account for them under one downstream protocol.

We test these questions with one downstream protocol: we hold the transformer body, subject splits, optimization and model-selection schedule fixed, and vary labeled subject count on two event-related potential (ERP) paradigms, P300 and N170, and on sleep-segment classification. The central comparison trains from scratch; separate masked-pretraining experiments test transfer under native and common prediction targets. Token counts and floating-point costs differ across representations, so these are matched training schedules rather than matched compute budgets. Further controls

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.

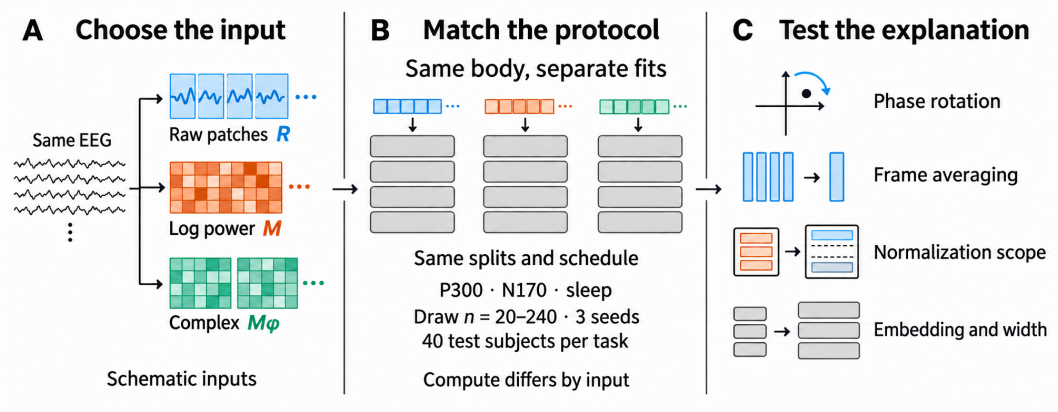

Figure 1: **Input choice changes the features available to a common learning protocol.** (A) Raw patches, log power, and complex coefficients transform the same EEG. (B) Separate models share the transformer architecture, subject splits, and training schedule, but not token counts or computation; n includes validation subjects. (C) Targeted comparisons test rotation, averaging, normalization scope, and embedding explanations. The artwork is schematic, not measured data; the completed study uses the inputs and geometry in Table 1.

vary normalization scope and the spectral embedding while retaining the corresponding input tokens (Figure 1).

The leading input depends on task and labeled-data scale. Raw patches outperform the fixed log-power input throughout the tested ERP range, contrary to our registered prediction, while log power helps sleep-segment classification at 20 and 40 subjects even after matching token counts and per-example compute, and its advantage diminishes as subjects are added. Within the tested Morlet pipeline, retaining complex coefficients substantially reduces the log-power deficit. Content-derived rotations by trial and frequency produce a large accuracy loss while preserving coefficient magnitudes before z-scoring. That intervention measures sensitivity to the implemented rotation; it changes cross-frequency relationships as well as alignment to the event. Learned envelope filters also improve over the fixed envelope, particularly on P300, so phase retention does not exhaust the useful choices.

The convolutional controls bound the architecture explanation to the designs we tested. Native-target pretraining gives raw a larger gain over scratch, but the arms predict different targets. With a common target, the experiment establishes neither a directional advantage nor equivalence between raw and complex-input gains.

2 Related Work

2.1 Input Choices and Their Ablations

The location of a spectral operation determines what an ablation tests. BIOT embeds Fourier energy vectors (Yang et al., 2023), while REVE and EEGPT learn from raw waveform patches (Ouahidi et al., 2025; Wang et al., 2024a). CBraMod combines a temporal convolutional branch with a Fourier-energy branch (Wang et al., 2024b). Its Appendix I compares temporal-only, frequency-only, and combined inputs on emotion and motor-imagery tasks, with the combined model leading: direct evidence from branch removal, not evidence that input representations have gone untested.

TFM-Tokenizer is a particularly close precedent: its input ablation compares raw-only, STFT-only, and combined temporal-spectral encoders, with the single-branch variants doubling their embedding dimensions to match the combined codebook width (Pradeepkumar et al., 2026). Its rankings depend on the metric: on TUEV, raw exceeds STFT in balanced accuracy, whereas STFT exceeds raw in Cohen's kappa. Matching codebook width does not establish equal parameter count or compute cost. We extend this line with subject-count curves and interventions within a fixed filterbank under a common supervised backbone.

Table 1: The study matches the training protocol, not computation. The top block describes fixed evaluation populations (E01); P300 and N170 share 18 people, giving 62 unique ERP evaluation people. The lower block uses 29-channel ERP geometry; MFLOPs count forward computation. All tasks use nested draws at $n \in \{20, 40, 80, 160, 240\}$ and three seeds; n includes the 15% subject-level validation split (S2).

Task	Classes	Channels	Eval. people	Eval. epochs
P300	2	29	40	8277
N170	2	29	40	6517
Sleep (1 s)	5	6	40	11998

Input	Frames/channel	Tokens	Parameters	MFLOPs
R	10	290	801,026	630
M	25	725	803,458	2221
M-phi	25	725	806,530	2226

2.2 Transform and target comparisons

BrainBERT compares STFT and superlet spectrograms for intracranial recordings (Wang et al., 2023; Moca et al., 2021). It maps each spectral column linearly into the transformer rather than through a two-dimensional convolutional embedding, and its transform ranking reverses between fine-tuned and frozen evaluation, showing why an input comparison must specify its adaptation regime. This is a spectral-transform comparison, distinct from replacing a spectral input with raw voltage.

LaBraM compares discrete spectral-token prediction with raw and continuous spectral reconstruction targets while retaining a raw-input encoder (Jiang et al., 2024b). Its raw-reconstruction variant leads on abnormality detection and trails the spectral-token variant on event classification, which concerns the training target and discretization rather than removal of phase from the encoder input. NeuroLM also places spectral supervision in its tokenizer (Jiang et al., 2024a). A raw path, a summed spectral branch, or a vector-quantized intermediate does not by itself establish invertibility or complete phase loss; Appendix S6 records observable input and embedding properties without assigning those unmeasured guarantees.

2.3 Phase and data scale

Temporal waveforms, time-frequency features, and phase dynamics already have demonstrated complementary value in ERP analysis (Ouyang & Zhou, 2023), and speech studies motivate front-end comparisons across training-set size (Palaz et al., 2013; Sainath et al., 2013; Tüske et al., 2014; Zeghidour et al., 2017). The new question here is the cost of specific front-end operations across labeled-data scales, not whether phase can carry useful neural information. Our phase-rotation contrast tests a concrete perturbation of complex coefficients, while normalization and convolutional controls test two competing explanations for the observed raw-versus-log-power deficit. The literature catalog distinguishes input replacement, branch removal, spectral-transform comparison, and prediction-target ablation; it does not support an exhaustive count of papers that have run a relevant control.

3 Methods

3.1 Shared training protocol

We held the transformer body, corpus within paradigm, subject splits, training epochs, optimizer, and learning-rate selection schedule fixed across scratch arms. Matching applies to the transformer body; the minimal input projection and positional table depend on token geometry. The transformer body (Vaswani et al., 2017) has four layers, width 128, and four attention heads. Total parameter count ranges from 801,026 for raw patches to 806,530 for complex coefficients, a spread below one percent. Table 1 makes the data and compute contract explicit.

The primary grid does not match compute: spectral inputs cost roughly 3.5 times as much per example on ERP and 2.8 times on sleep (S2.1). An additional sleep block matches tokens and per-example

compute in both directions: log power at 10 frames per channel versus raw 100 ms patches, and log power at 25 frames versus raw 40 ms patches (C13-T). The matched contrasts also change spectral averaging resolution or raw patch length. All four arms are rerun within the block (S2.1, S5.6). We used AdamW (Loshchilov & Hutter, 2019), cosine decay without restarts (Loshchilov & Hutter, 2017), batch size 64, and 30 epochs, with each arm selecting its learning rate from {3e-4, 1e-3} and its best epoch on validation data. This two-point grid supports a common search schedule, not a claim that each arm's optimal learning rate was bracketed.

3.2 Data and splits

PURSUE supplies P300 target/standard and N170 face/car tasks (Couperus et al., 2026b;a), extending ERP CORE (Kappenman et al., 2020). Sleep uses five PhysioNet/CinC Challenge 2018 classes (Ghassemi et al., 2018). Each paradigm had a fixed evaluation population of 40 held-out subjects. Training draws at {20, 40, 80, 160, 240} subjects were nested within seed and varied across three seeds; evaluation subjects entered no training, validation, or pretraining pool. P300 and N170 evaluation populations partly overlap by person, a dependence accounted for in the amended statistical specification (S3).

Preprocessing resampled to 200 Hz, applied a zero-phase 0.5–45 Hz band-pass and average reference, and retained one-second epochs without artifact rejection. ERP epochs spanned -200 to +800 ms relative to the stimulus. Sleep also used one-second windows to preserve sequence geometry across paradigms, rather than the conventional 30-second staging window, so absolute sleep scores are not comparable to standard staging benchmarks. Sleep uses six channels and ERP 29; a separate six-electrode P300 control tests the channel-count explanation. S2 gives draw, validation-split, and epoch-sampling details.

3.3 Input representations

R forms 100 ms raw patches after per-recording robust scaling (Perslev et al., 2021). M applies a fixed Morlet filterbank with 24 log-spaced centers from 2 to 42 Hz (Cohen, 2019), computes log10 power at each 200 Hz frame, averages those log-power frames to 25 Hz, and z-scores per frequency using training statistics. M-phi retains real and imaginary coefficients from the same filterbank and subsamples them, yielding 48 coordinates per channel and frame; both spectral outputs have 25 frames per second. Complex coefficients retain explicit phase coordinates, but finite bandwidth and subsampling mean this does not establish lossless waveform recovery. L-band and L-random use sinc filters with trainable cutoffs (Ravanelli & Bengio, 2018; Zeghidour et al., 2021), initialized on the Morlet grid or randomly, followed by power averaging and log compression. Unlike M, L pools power before taking the logarithm; filter shape and normalization also differ, so their comparison does not isolate learnability alone.

Raw, spectral, and learned arms initially follow different normalization conventions: robust per-recording scaling, pooled per-frequency z-scoring, and batch normalization. Additional raw-pooled and spectral-per-recording arms match normalization scope in both directions, keeping the two statistics distinct (S5.5).

3.4 Phase-rotation and averaging contrasts

M-sub replaces M's frame averaging with subsampling at the same 25 Hz output rate. M-phi-scram multiplies each trial's complex coefficients at frequency f by a unit complex rotation with one content-derived angle per trial and frequency, shared across time and channels, and fixed for each epoch across training seeds. Before arm-specific coordinate z-scoring, that rotation preserves coefficient magnitudes and within-frequency temporal and interchannel phase differences, while changing cross-frequency relationships and phase alignment across trials.

We compare M-sub minus M (averaging), M-phi minus M-phi-scram (rotation), and M-phi-scram minus M-sub (the remaining representation contrast, which changes coordinates, nonlinear accessibility, width, and nuisance structure together). The registered reading asked whether the rotation term was largest. An averaged-complex arm was excluded because averaging complex coefficients would itself filter the analytic signal, so this is a partial decomposition rather than a complete factorial (S2.3).

3.5 Embedding and pretraining controls

Linear spectral-column embeddings can absorb a fixed permutation into their weights, making that permutation an invalid ordering control. We therefore compared log-power, complex, and frequency-permuted log-power inputs under a two-dimensional convolutional embedding, leaving the input tensors, token counts, transformer body, and classifier unchanged. A wider convolutional variant tests the parameter-budget objection (S6). These tests concern one embedding family; convolution alone does not implement a squaring operation or guarantee improved energy extraction.

Self-supervised learning (SSL) is evaluated by the contrast (SSL minus scratch) for M-phi minus the same difference for R on the ERP tasks at 20 and 80 subjects. The native-token block fine-tunes one pretrained model per arm at six training-draw seeds under deterministic algorithms; its seed interval carries a device component, and the arms reconstruct different targets. The common-target block masks 200 ms waveform blocks before each front end and reconstructs the same robust-scaled raw samples. It uses six pretraining seeds per arm and reruns scratch comparators within the block; its six-seed interval also carries pretraining, node and device components. The blocks differ in masking geometry and variance structure, so their comparison is descriptive. S2.8 and S5.3 specify search-budget, wavelet-context and native-target limitations and the manipulation checks.

3.6 Outcomes and uncertainty

Balanced accuracy is computed from pooled evaluation trials within each fitted model, then averaged equally over seeds and, for the ERP composite, paradigms. The primary gap is G(n) = balanced accuracy(M) minus balanced accuracy(R). The original reading required a positive gap with a 95 percent interval excluding zero at n in {20, 40, 80}. The archive, original analysis, and post-review dependence correction are distinguished in S3. Corrected intervals describe evaluation-person uncertainty conditional on the fitted models, selected training draws, and observed cohort membership. They do not quantify training randomness or changing cohort composition; per-seed spreads are reported separately. Every interval in this paper is pointwise: none is a simultaneous band over subject counts, arms, or claims, and no reported adjustment provides familywise coverage of the study (S3.3). Published-checkpoint probes fit ridge readouts of Morlet log power against matched random-initialization nulls (S2.9). S7 maps the reported comparisons to visible evidence identifiers and marks missing provenance.

4 Results

4.1 Raw outperforms log power across the tested ERP scales

The ERP gap G(n) = balanced accuracy(M) minus balanced accuracy(R) is negative throughout the subject-count grid: -0.0925 at n = 20, -0.1126 at n = 80, and -0.1122 at n = 240 (Figure 2A; C1, S1). The separate gaps range from -0.0462 to -0.0835 on P300 and -0.1173 to -0.1423 on N170. The fixed log-power input therefore fails the registered raw-versus-spectral prediction in both evoked paradigms.

Matching normalization scope in both directions leaves ERP deficits of approximately 0.09–0.12 balanced accuracy (C19, Figure 3A, S5.5). The raw-pooled and spectral-per-recording controls therefore do not explain away the raw advantage. Their paired changes across subject count do not establish that the deficit deepens with data: the observation is persistence over the tested range, not a claim that more data can never help.

4.2 Complex coefficients recover much of the deficit; rotation removes the gain

Keeping complex Morlet coefficients improves on log power at all ten ERP cells, by +0.0245 to +0.1344 (C11), recovering 53–101 percent of the raw-minus-log-power gap with a median of 72 percent. Raw still leads at nine of ten cells; N170 at n = 40 is the exception. No equivalence margin was registered, so recovery does not establish equivalence to raw.

Learned envelope filters also improve on the fixed envelope, by +0.0195 to +0.0821 over twenty comparisons. On N170 they remain below complex coefficients throughout; on P300 they often

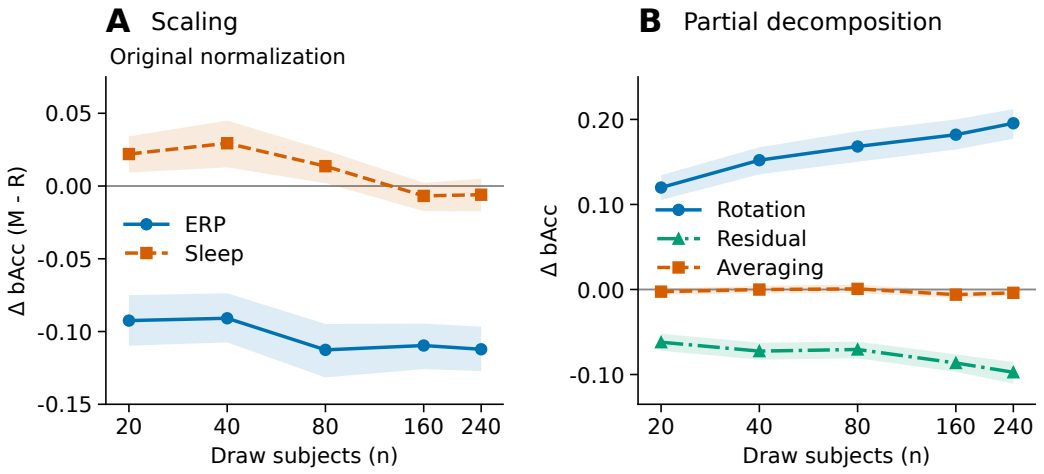

Figure 2: **The ERP log-power deficit persists, and rotation dominates the partial decomposition.** (A) Balanced accuracy(M) minus balanced accuracy(R) for the ERP composite and one-second sleep segments under the original normalization conventions. (B) ERP contrasts: rotation, M-phi minus M-phi-scram; residual, M-phi-scram minus M-sub; averaging, M-sub minus M. Points average three fitted seeds and, for ERP, two paradigms; bands are pointwise conditional 95 percent evaluation-person intervals, not retraining uncertainty. Each null is zero difference. Rotation preserves magnitudes and within-frequency phase differences before coordinate z-scoring but changes event alignment and cross-frequency relationships. The residual bundles coordinates, width, and accessibility; averaging and subsampling share a 25 Hz output rate (C1, C12, C13; E01).

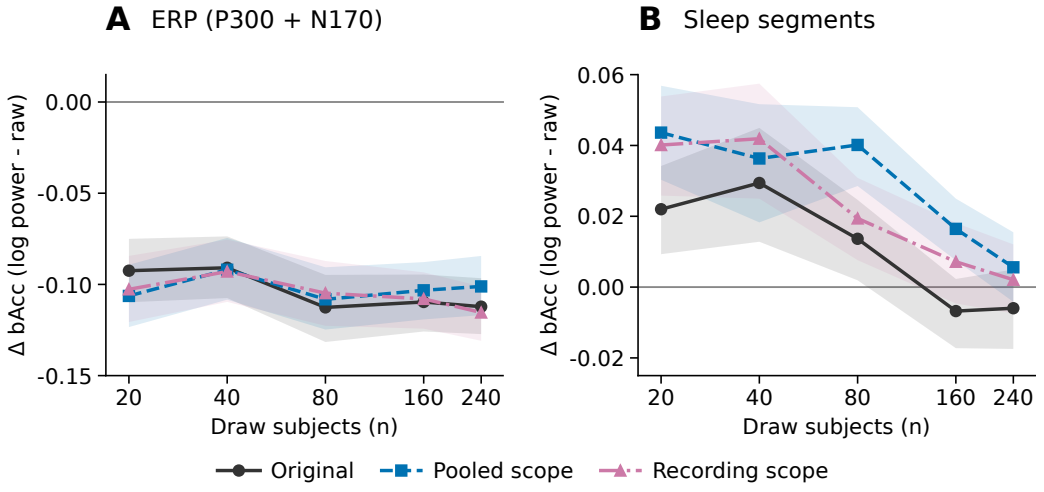

Figure 3: **Matching normalization scope retains the ERP deficit but changes the sleep endpoint.** (A) ERP composite. (B) One-second sleep segments. Curves compare M minus R (original), M minus pooled-normalized raw (pooled scope), and recording-normalized waveform followed by M minus R (recording scope). Negative values favor raw; zero is each contrast's null. Points average three fitted seeds and, for ERP, two paradigms; bands are pointwise conditional 95 percent evaluation-person intervals. Scope matching retains different statistics: robust waveform scaling versus pooled spectral z-scoring. Both matched sleep endpoint estimates are positive but their intervals include zero (C19; E04, S5.5).

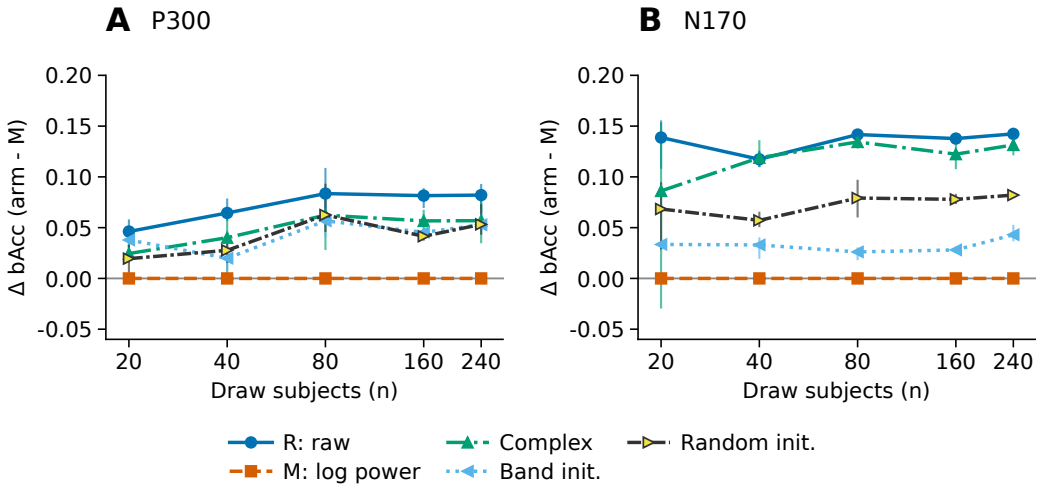

Figure 4: **Learned envelopes narrow the gap with a task-dependent ordering.** (A) P300. (B) N170. Each point is the mean paired-seed accuracy gain over fixed log power M; the zero line is M itself. Vertical bars span the minimum and maximum of three paired-seed gains, not a confidence interval; all fits remain, including the collapsed N170 complex fit at n = 20. Intervals elsewhere in this paper are pointwise, not simultaneous (S3.3). R is raw, M-phi is complex, and L-band/L-random are learned envelopes with different initializations. Their filter shape and normalization also differ from M, so this does not isolate learnability. The full absolute-accuracy ladder appears in S6 (C11; E01).

approach that arm, with exceptions at n = 20 and n = 80 (Figure 4; the complete absolute-accuracy ladder is in S6). Both filterbank choice and coefficient representation matter, and the clean phase-based grouping applies more strongly to N170.

The registered decomposition makes the larger intervention explicit. Removing the per-trial, per-frequency rotation raises ERP balanced accuracy by +0.1199 at n = 20 and +0.1955 at n = 240, the largest term at every subject count (C12; Figure 2B). Frame averaging changes accuracy by at most about 0.006 in magnitude, at the same output sampling rate; its n = 240 interval excludes zero negatively, without approaching the rotation effect's magnitude. The remaining rotated-complex-minus-subsampled-log-power contrast is negative, from -0.0620 to -0.0974.

The rotation result demonstrates a large dependence on this intervention and its downstream normalization effects. It does not isolate an exclusively event-locked neural mechanism, because cross-frequency relationships change too, and the negative residual does not measure a dimensionality penalty, because the compared arms expose different coordinates and nonlinear features to the embedding.

4.3 The sleep advantage diminishes as labeled data grows

Log power leads raw on sleep at small subject counts: G(n) is +0.0220 at n = 20 and +0.0136 at n = 80, reaching -0.0060 at n = 240 under the original normalization conventions (C13). The relevant result is the diminishing advantage; the small negative endpoint does not establish a reliable late raw advantage, and the matched-scope endpoints, +0.0056 (pooled) and +0.0021 (recording), have conditional intervals including zero (Figure 3B). A crossover is therefore not robust to the normalization control.

The sleep gap exceeds the ERP composite by 0.1028–0.1263 across the grid, and complex coefficients rank differently there: M-phi trails M on sleep from -0.0364 at n = 20 to -0.0066 at n = 240. This is a task interaction in the observed representation costs, not evidence that phase is universally useful or universally nuisance on sleep. The six-electrode P300 control retains a negative gap, -0.0464 to

Table 2: Convolution uses frequency order but does not remove the ERP deficit. Cells give balanced-accuracy differences and conditional 95% person intervals at the three registered subject counts. Zero is the named arm contrast's null. Conv. denotes convolutional embedding; the input contrast uses it on both arms. ERP points average two tasks and three fitted seeds; intervals condition on fits, draws and cohort membership. Both complex-convolution comparisons are required by C16; failure to exclude zero positively is not equivalence. Intervals are pointwise, not simultaneous across contrasts or subject counts (S3.3). S6 reports all five counts, seed spreads and collapsed fits (E02, E03).

Comparison	$n = 20$	$n = 40$	$n = 80$
Log-power conv. − raw	-0.0881 [-0.1073, -0.0696]	-0.0964 [-0.1138, -0.0781]	-0.1099 [-0.1290, -0.0910]
Ordered − permuted conv.	+0.0131 [+0.0072, +0.0187]	+0.0113 [+0.0038, +0.0189]	+0.0065 [+0.0009, +0.0123]
Complex − log power (conv.)	+0.0482 [+0.0364, +0.0597]	+0.0426 [+0.0313, +0.0539]	+0.0938 [+0.0809, +0.1071]
Conv. − linear (complex)	-0.0027 [-0.0092, +0.0039]	-0.0424 [-0.0498, -0.0347]	-0.0018 [-0.0091, +0.0063]
Wide − linear (complex)	-0.0010 [-0.0081, +0.0059]	-0.0048 [-0.0110, +0.0014]	-0.0100 [-0.0168, -0.0032]

-0.0663 compared with -0.0462 to -0.0835 on the full montage (S5.1), so channel count alone does not account for the regime difference.

Matching token count and per-example compute in both directions retains the sleep advantage at n = 20 and 40 (C13-T; E14, S5.6): the matched gaps are +0.0242 and +0.0216 at 20 subjects, and +0.0280 and +0.0197 at 40. Both directions retain at least half their block's own unmatched gap, meeting the registered point-estimate reading, SURVIVES, at those two counts. The registration assigns no verdict at the other three counts; their contrasts are descriptive. Neither a robust crossover nor a broader sleep prescription follows. The conditional intervals do not include seed variation, which exceeds the mean gap in two n = 20 comparisons.

4.4 Architecture controls and the bounded pretraining contrast

Table 2 exposes the registered architectural tests. A convolutional spectral embedding leaves the log-power deficit against raw at -0.0881 at n = 20 and -0.1083 at n = 240 (C15, S6). Frequency permutation reduces the convolutional arm's performance, showing that this embedding uses frequency order even though it does not rescue the log-power input. The complex-input convolution improves on its log-power counterpart but fails the registered requirement to beat both that counterpart and the linear complex-input arm (C16), and a wider convolution does not restore the predicted advantage (C18). This is a failed prediction about the tested embedding family. It supplies no evidence that an unoccupied category in a literature map would be an optimal architecture.

Native-target pretraining gives raw a larger gain over scratch: the M-phi-minus-R gain contrast is -0.021068, with a six-seed 95 percent t interval of [-0.0396, -0.0025], a seed interval carrying a device component (C4-E3; E12). The arms reconstruct different targets, so this does not isolate phase. Under the common waveform target, it is +0.008896 with a 95 percent t interval of [-0.013926, +0.031719] over six paired pretraining and fine-tuning seeds, carrying node and device components (C4-E4; E16, S5.3). The registered reading is E4-NULL: neither direction nor equivalence is established. E3 and E4 differ in objective, masking geometry and variance structure; their descriptive comparison does not explain the native result.

Published raw-input encoders can nonetheless represent spectral features: REVE's ridge readout predicts Morlet log power at a delta of +0.446 over its matched random-initialization null, and LaBraM's corresponding delta is +0.061 (C5, C6; S2.9; (Ouahidi et al., 2025; Jiang et al., 2024b)). These probes establish checkpoint-specific linear accessibility, not an information ceiling or a general account of pretraining.

5 DISCUSSION

On the ERP tasks, complex coefficients recover much of the log-power deficit. Sleep gives a different ordering: log power leads at 20 and 40 subjects after token and per-example compute matching, then loses ground as subjects are added. The matched endpoints do not establish a reliable crossover. Normalization scope also changes the estimates, so a reported ranking should state how each input was scaled.

The phase intervention supports a more specific interpretation than a raw-versus-spectral ranking. Coefficient magnitudes stay fixed before arm-specific z-scoring, so the contrast originates in the rotation and its downstream normalization effects. Event alignment is one such property; cross-frequency relationships are another, and the experiment does not separate their contributions. Likewise, the small averaging effect concerns two constructions at the same output rate, and the residual contrast bundles token width with coordinate and accessibility changes. The content-derived angles add another limit: a small registered P300 diagnostic found no detectable class difference in resultant lengths, but did not establish label independence or validate the production angles (S2.3). The large intervention effect does not identify an exclusive neural mechanism.

Magnitude summaries expose energy while removing explicit phase coordinates. In the REVE checkpoint, a linear probe predicts Morlet log power above its matched random-initialization null. That probe does not establish lossless encoding or an information ceiling. The post-hoc accessibility predictor failed (C14), leaving no validated rule for selecting inputs from event locking or sample count.

Frequency-structured convolution does not remove the ERP log-power deficit. Complex-input convolution does not beat both neighboring arms, even with a wider embedding. We withdrew the original design prescription after finding that its literature map mislabeled inputs as lossless and BrainBERT as convolutional (S6).

The transformer is small, the learning-rate grid is coarse, and equal epochs give spectral arms more computation per example in the primary grid. Several complex-input fits also collapse at small subject counts; a wider grid did not rescue them. The normalization controls test scope in two directions but do not equate normalization statistics or cover learned-arm batch normalization. The one-second sleep window limits comparison with conventional staging, and conditional evaluation-person intervals omit uncertainty from new training draws, seeds, and changing cohort composition. Transfer remains unresolved. Raw gains more under native targets, which differ across arms; the common-target experiment supports neither a direction nor equivalence for complex versus raw gains. Its masking changes wavelet context, and its seed interval carries node and device components (S5.3). For future comparisons, choose inputs for the intended task and labeled-data budget, and report phase treatment, normalization, token geometry, and prediction target.

6 CONCLUSION

Raw patches led log power on both ERP tasks. On one-second sleep segments, log power led at 20 and 40 subjects after both token and compute matching, but that advantage diminished at larger counts. Complex coefficients recovered much of the ERP gap. Phase rotation produced the largest change, but it also changed cross-frequency relationships and event alignment. The common-target interval includes zero, so it establishes neither a directional advantage nor equivalence between raw and complex-input gains. Future comparisons should state the task, labeled-data budget, phase treatment, normalization, token geometry, and prediction target.

7 REPRODUCIBILITY STATEMENT

Run records, subject manifests, and code are retained in a private evidence archive. The revised analysis corrects evaluation-person dependence while preserving the original measurements and reports; its amendment is explicitly post-review. A code-only anonymous snapshot has been prepared and checked locally; it has not yet been uploaded as a public repository. The transfer contrast is reported from a deterministic re-run whose same-device repeats agree exactly; why earlier reruns drifted is not explained (S5.3, S7). Section S7 states the archive, analysis, and availability boundaries.

8 Ethics Statement

This study uses three publicly available human electroencephalography corpora, all previously released for research use by their originating projects under their own consent and ethics approvals; we did not collect new human data and had no contact with participants.

9 Use of Large Language Models

We used a large-language-model coding assistant throughout this project and are responsible for everything in this paper, including anything it contributed. It wrote and edited much of the experiment, analysis, figure, and test code under our direction; reviewed experimental designs before they ran, including the normalization-scope control and its pre-registered reading; and drafted and revised prose throughout, including interpretive framing. The hypotheses, the pre-registered readings, and the decisions about what each result licenses are ours.

References

Michael X Cohen. A better way to define and describe morlet wavelets for time-frequency analysis. *NeuroImage*, 199:81–86, 2019. doi: 10.1016/j.neuroimage.2019.05.048. URL https://doi.org/10.1016/j.neuroimage.2019.05.048.

J. W. Couperus, C. M. Bukach, and C. L. Reed. PURSUE N170 face perception. NEMAR data set, 2026a. URL https://nemar.org/dataset/on007096. Accession citation; the historical acquisition snapshot used by our experiments is not verified.

J. W. Couperus, C. M. Bukach, and C. L. Reed. PURSUE P300 visual oddball. NEMAR data set, 2026b. URL https://nemar.org/dataset/on007056. Accession citation; the historical acquisition snapshot used by our experiments is not verified.

Arnaud Delorme, Dung Truong, Choonhan Youn, Subhashini Sivagnanam, Claire Stirm, Kenneth Yoshimoto, Russell A. Poldrack, Amitava Majumdar, and Scott Makeig. NEMAR: an open access data, tools and compute resource operating on neuroelectromagnetic data. *Database*, 2022:baac096, 2022. doi: 10.1093/database/baac096. URL https://academic.oup.com/database/article/doi/10.1093/database/baac096/6823529.

Mohammad M. Ghassemi, Benjamin E. Moody, Li-wei H. Lehman, Christopher Song, Qiao Li, Haoqi Sun, Roger G. Mark, M. Brandon Westover, and Gari D. Clifford. You snooze, you win: The PhysioNet/Computing in Cardiology challenge 2018. In *Computing in Cardiology*, volume 45, pp. 1–4, 2018. doi: 10.22489/CinC.2018.049. URL https://physionet.org/content/challenge-2018/1.0.0/. Primary challenge description verified 2026-09-12. The cited release page identifies the source; it does not establish the version used by our archived runs.

Wei-Bang Jiang, Yansen Wang, Bao-Liang Lu, and Dongsheng Li. Neurolm: A universal multi-task foundation model for bridging the gap between language and eeg signals. *The Thirteenth International Conference on Learning Representations, 2025*, 2024a. URL https://arxiv.org/abs/2409.00101.

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

Emily S. Kappenman, Jaclyn L. Farrens, Wendy Zhang, Andrew X Stewart, and Steven J. Luck. Erp core: An open resource for human event-related potential research. *NeuroImage*, 2020. doi: 10.1016/j.neuroimage.2020.117465. URL https://doi.org/10.1016/j.neuroimage.2020.117465.

Vernon J. Lawhern, Amelia J. Solon, Nicholas R. Waytowich, Stephen M. Gordon, Chou P. Hung, and Brent J. Lance. Eegnet: A compact convolutional network for eeg-based brain-computer interfaces. *arXiv:cs.LG*, 2016. doi: 10.1088/1741-2552/aace8c. URL https://arxiv.org/abs/1611.08024.

Ilya Loshchilov and Frank Hutter. SGDR: Stochastic gradient descent with warm restarts. In *International Conference on Learning Representations*, 2017. URL https://arxiv.org/abs/1608.03983v5.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://arxiv.org/abs/1711.05101v3.

Vasile V. Moca, Harald Barzan, Adriana Nagy-Dabacan, and Raul C. Muresan. Time-frequency super-resolution with superlets. *Nature Communications*, 12:337, 2021. doi: 10.1038/s41467-020-20539-9. URL https://doi.org/10.1038/s41467-020-20539-9.

Yassine El Ouahidi, Jonathan Lys, Philipp Thölke, Nicolas Farrugia, Bastien Pasdeloup, Vincent Gripon, Karim Jerbi, and Giulia Lioi. Reve: A foundation model for eeg – adapting to any setup with large-scale pretraining on 25,000 subjects. *arXiv:cs.LG*, 2025. URL https://arxiv.org/abs/2510.21585.

Guang Ouyang and Changsong Zhou. Exploiting information in event-related brain potentials from average temporal waveform, time–frequency representation, and phase dynamics. *Bioengineering*, 10(9):1054, 2023. doi: 10.3390/bioengineering10091054. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC10525145/.

Dimitri Palaz, R. Collobert, and M. Magimai-Doss. Estimating phoneme class conditional probabilities from raw speech signal using convolutional neural networks. In *Interspeech*, pp. 1766–1770, 2013.

Mathias Perslev, S. Darkner, Lykke Kempfner, M. Nikolic, P. Jennum, and C. Igel. U-sleep: resilient high-frequency sleep staging. *NPJ Digital Medicine*, 4, 2021.

Jathurshan Pradeepkumar, Xihao Piao, Zheng Chen, and Jimeng Sun. Tokenizing single-channel eeg with time-frequency motif learning. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/3dc85735f6e2fcf093e67b134fa00d21-Abstract-Conference.html. Input ablations checked against arXiv:2502.16060v3, Sections 4.5, C.2 and D.1, accessed 2026-09-12.

Mirco Ravanelli and Yoshua Bengio. Speaker recognition from raw waveform with sincnet. *arXiv:eess.AS*, 2018. URL https://arxiv.org/abs/1808.00158.

Tara N. Sainath, Brian Kingsbury, Abdel rahman Mohamed, and B. Ramabhadran. Learning filter banks within a deep neural network framework. *2013 IEEE Workshop on Automatic Speech Recognition and Understanding*, pp. 297–302, 2013.

Robin Tibor Schirrmeister, Jost Tobias Springenberg, Lukas Dominique Josef Fiederer, Martin Glasstetter, Katharina Eggensperger, Michael Tangermann, Frank Hutter, Wolfram Burgard, and Tonio Ball. Deep learning with convolutional neural networks for EEG decoding and visualization. *Human Brain Mapping*, 38(11):5391–5420, 2017. doi: 10.1002/hbm.23730. URL https://arxiv.org/abs/1703.05051.

Jamiyan Sukhbaatar, Satoshi Imamura, and Toshihisa Tanaka. Bandvq: Band-wise vector-quantized eeg foundation model. *arXiv:cs.LG*, 2026. URL https://arxiv.org/abs/2605.24921.

Zoltán Tüske, Pavel Golik, Ralf Schlüter, and H. Ney. Acoustic modeling with deep neural networks using raw time signal for lvcsr. In *Interspeech*, pp. 890–894, 2014.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://papers.nips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

Christopher Wang, Vighnesh Subramaniam, Adam Uri Yaari, Gabriel Kreiman, Boris Katz, Ignacio Cases, and Andrei Barbu. Brainbert: Self-supervised representation learning for intracranial recordings. *arXiv:cs.LG*, 2023. URL https://arxiv.org/abs/2302.14367.

Guangyu Wang, Wenchao Liu, Yuhong He, Cong Xu, Lin Ma, and Haifeng Li. Eegpt: Pretrained transformer for universal and reliable representation of eeg signals. In *Advances in Neural Information Processing Systems 37 (NeurIPS 2024)*, 2024a. doi: 10.52202/079017-1239. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/4540d267eeec4e5dbd9dae9448f0b739-Abstract-Conference.html.

Jiquan Wang, Sha Zhao, Zhiling Luo, Yangxuan Zhou, Haiteng Jiang, Shijian Li, Tao Li, and Gang Pan. Cbramod: A criss-cross brain foundation model for eeg decoding. *arXiv:eess.SP*, 2024b. URL https://arxiv.org/abs/2412.07236.

Chaoqi Yang, M. Brandon Westover, and Jimeng Sun. Biot: Cross-data biosignal learning in the wild. *arXiv:eess.SP*, 2023. URL https://arxiv.org/abs/2305.10351.

Zhizhang Yuan, Fanqi Shen, Meng Li, Yuguo Yu, Fei Wu, Chenhao Tan, and Yang Yang. Brainwave: A brain signal foundation model for clinical applications. *arXiv:q-bio.NC*, 2024. URL https://arxiv.org/abs/2402.10251. Earlier arXiv revisions titled "Brant-2: Foundation Model for Brain Signals".

Neil Zeghidour, Nicolas Usunier, Iasonas Kokkinos, Thomas Schatz, Gabriel Synnaeve, and Emmanuel Dupoux. Learning filterbanks from raw speech for phone recognition. *arXiv:cs.CL*, 2017. URL https://arxiv.org/abs/1711.01161.

Neil Zeghidour, Olivier Teboul, Felix de Chaumont Quitry, and Marco Tagliasacchi. LEAF: A learnable frontend for audio classification. In *International Conference on Learning Representations (ICLR)*, 2021. URL https://arxiv.org/abs/2101.08596.

Daoze Zhang, Zhizhang Yuan, Yang Yang, Junru Chen, Jingjing Wang, and Yafeng Li. Brant: Foundation model for intracranial neural signal. *Advances in Neural Information Processing Systems 36*, 2023. doi: 10.52202/075280-1144. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/535915d26859036410b0533804cee788-Abstract-Conference.html.

A S1. Registered readings and corrections

A.1 S1.1 Claim status

The claim ledger records each hypothesis, observable, null, decision rule, and status. Entries made after the corresponding result existed are exploratory. The table summarizes original readings; S3 details the post-review dependence correction and S7 indexes the evidence. A hypothesis can be unsupported while its opposite measured contrast remains informative.

	claim	status
C1	headline front-end gap, the central claim	**not supported**
C2	the gap is spectral structure, not denoising	inconclusive, instrument invalid
C3	learned front ends converge toward the fixed grid	**not supported**
C4	self-supervised delta per front end	original hypothesis **not supported**; raw gained more under native objectives
C5	a raw-input checkpoint linearly encodes the spectrum	supported
C6	a second checkpoint barely encodes it beyond its input stage	supported
C7	published checkpoints are band-selective at the input stage	inconclusive

	claim	status
C8	input stages are tuned to high frequencies	exploratory, rephrased
C9	raw pretraining fails to recover spectral structure	**withdrawn**
C10	the headline verdict does not depend on the compute stack	**not supported**
C11	the failing ingredient is the log-power convention	supported, in part
C12	rotation, residual and averaging contrast ordering	supported; no unique neural mechanism
C13	the front-end answer is regime-dependent	supported
C14	linear accessibility predicts the trained ordering	**refuted as written**
C15	the verdict under a frequency-structured architecture	supported
C16	the empty cell of the design map	**not supported**
C17	whether the convolution buys ordering or only capacity	supported (ordered versus permuted convolution)
C18	whether C16's architecture axis is capacity-limited	supported (confound ruled out)
C19	normalization scope versus representation	ERP deficit retained; deepening not supported

The later registered controls retain their own readings: C13-T is supported as SURVIVES at 20 and 40 subjects under both token/compute matches (E14); P2-A's narrow angle-label diagnostic is inconclusive (E15); C4-E4 is inconclusive as E4-NULL and does not change C4's native-objective status (E16).

S6 gives the architecture results in detail: C17 supports frequency ordering over its permuted control but not an advantage of convolution over the linear embedding. C16's conjunction failed; its lossless/empty-category rationale was not established, and its associated design prescription was withdrawn (S6.3). C18 shows only that the tested capacity increase did not restore the predicted advantage; it does not identify a unique cause.

A.2 S1.2 Original hypotheses and failed controls

C1, C3, C4 and C10 each failed their registered reading; S4.2 states the measured values that replaced them and is not repeated here. C2's frequency permutation is algebraically uninformative under the linear spectral embedding, so its proposed structure-versus-denoising interpretation is void (S4.1).

C14 was written after the relevant trained results already existed, despite initially being described as predictive. Its own refutation condition also fired: weak linear accessibility did not prevent the raw arm from reaching the best trained sleep performance at the larger subject counts. It remains exploratory and refuted as written; the paper claims no training-free front-end selection rule.

A.3 S1.3 Corrections and available evidence

Correction	Current evidence or boundary
Ridge fits without an intercept produced R-squared values -2.76 and -4.19. C9 was withdrawn.	Corrected checkpoint-versus-random probes supersede those values (S2.9, E07).
A driver change split a sweep across software stacks.	Cross-stack point variability is retained; the old containment analysis is historical (S4.2, E09).

Correction	Current evidence or boundary
C14 was entered after its target result existed.	Exploratory status and the failed reading remain recorded.
A stale model-size label, an obsolete learning-rate audit and an over-rounded permutation bound entered draft prose.	Code-derived counts, current R/P300 grid audit and explicit comparison to seed spread replace them.
A pooled activation-scale measurement had no retained output, and an early failed convolution run had been overwritten.	The scale diagnostic is reproducible; the original failed run remains unavailable (S4.4, E10).
Subject resampling did not preserve shared people across fitted seeds and ERP tasks.	Reviewed corrected reports preserve point estimates and replace active intervals (S3).
The literature map misclassified BrainBERT and omitted verified input comparisons.	Operational input/embedding categories replace the lossless/empty-cell prescription (S6.6).

Original readings and superseded reports remain preserved; the active analysis uses the reviewed dependence correction without claiming retroactive preregistration. The withdrawn C9 values do not trace to passing numeric checks, and unavailable early observations do not support the headline. The result, claim, numeric-provenance, and ledger checks have distinct scopes; none alone verifies scientific interpretation.

B S2. Full experimental detail

Main-text Methods carries what a reader needs to follow the contrast: the matching definition, the arms, the paradigms, the metric, and the reading fixed in advance. This section states the implemented protocol and the recoverable run details. S7 distinguishes preserved evidence from provenance gaps that prevent an exact rebuild.

Architecture counts in S2.1 are emitted from the code. Dataset, split and probe details below come from their respective configurations and archived records, indexed in S7.

B.1 S2.1 Per-arm parameters, token geometry, and floating-point cost

Size M (the headline size: 4 layers, model width 128, 4 heads), 29 channels, which is the evoked corpora after excluding electro-oculogram channels.

Arms with identical geometry share a row.

arm	token dim	frames	tokens	body params	embed params	total params	MFLOPs
R	20	10	290	798,338	2,688	801,026	630
M, M-sub, L-band, L-random, D2	24	25	725	800,258	3,200	803,458	2221
M-phi, M-phi-scram	48	25	725	800,258	6,272	806,530	2226
M-conv, D2-conv	24	25	725	800,258	4,352	804,610	2227
M-phi-conv	48	25	725	800,258	4,472	804,730	2231

Every primary cell records its own runtime stack, and the whole primary grid ran on one:

```
python  3.13.15          gpu     Tesla V100-SXM2-32GB
torch   2.8.0+cu128      driver  580.82.07
cuda    12.8             cudnn   91002
```

S5.2 measures how far a headline cell moves across deliberately different stacks; that comparison is between its measurement and this single-stack primary grid, not within it.

Cross-arm total-parameter spread across the arms sharing the linear embedding convention is **0.687 percent**, against the 2 percent tolerance asserted in code. The convolutional arms are excluded from that group by construction: they carry a different embedding, and what is asserted of them instead is that their transformer body is identical to their linear partner's.

The body is identical modulo token count. Arm R shows 798,338 against the spectral arms' 800,258; the difference, 1,920, is exactly the 15 fewer rows in its time positional-embedding table (10 frames against 25) times the model width of 128.

The primary grid does not match floating-point cost. The spectral arms cost about 3.5 times the raw arm per example on the 29-channel ERP montage, emitting 725 tokens against 290. On the six-channel sleep montage the corresponding ratio is 2.8, with 150 tokens against 60; the token ratio is 2.5 in both cases.

C13-T matches sleep tokens and per-example compute in both directions (E14). All four arms run under deterministic algorithms on one node, at the original five subject counts and three seeds, giving 60 cells. The unmatched R and M arms are rerun inside this block; no archived value enters a matched contrast.

Sleep arm	Frames/channel	Tokens	MFLOPs
R, 100 ms patches	10	60	102.1
M-10f, 100 ms averaging	10	60	102.1
M, 40 ms averaging	25	150	282.9
R-25f, 40 ms patches	25	150	282.3

Direction A compares M-10f with R; direction B compares M with R-25f. Matched-pair compute differs by at most 0.2 percent and the four-arm parameter spread is 0.30 percent. Each arm retains its base normalization convention, the two-point learning-rate grid, 30 epochs, batch size 64 and the original subject draws and evaluation population. Coarser spectral averaging changes temporal resolution; shorter raw patches change receptive field and embedding width while retaining every input sample. Thus the control does not vary token count alone. S5.6 reports its registered reading and complete curve.

The convolutional embedding is not a capacity increase. `M_conv` adds 1,152 parameters over `M`, and `M_phi_conv` has 1,800 **fewer** than `M_phi`, because a convolution shares weights across frequency where the linear embedding held a full 48-by-128 matrix. `D2_conv` is exactly parameter-matched to `M_conv`, which makes C17 a control for capacity rather than an approximation to one.

B.2 S2.2 Preprocessing

Identical across every from-scratch arm. Recordings were resampled to 200 Hz, band-pass filtered 0.5 to 45 Hz with a zero-phase finite impulse response filter, average referenced, and amplitude scaled by 100 microvolts so values fell mostly within [-1, 1], following the convention used by CBraMod and LaBraM.

No artifact rejection was applied. The corpus we surveyed reports a minimal-preprocessing optimum, and an artifact-suppressed variant existed only as a separate arm that is not reported here.

Epochs were stimulus-locked from -200 to +800 ms, giving 200 samples.

B.3 S2.3 FRONT-END SPECIFICATIONS

R (raw) formed per-channel 100 ms patches of 20 samples with a linear embedding, after per-recording robust scaling (median 0, interquartile range 1, clipped at 20 times the interquartile range).

M (log power) applied a full-width-half-maximum-parameterized Morlet family with 24 center frequencies log-spaced from 2 to 42 Hz, spectral width one third of the center frequency clamped to [1, 10] Hz. It computed base-10 log power at 200 Hz, averaged those values to 25 Hz, then z-scored per frequency using training-set statistics.

M-phi (phase-preserving) used the identical filterbank and emitted real and imaginary coefficients, 48 per input channel, z-scored per frequency and decimated by subsampling at the center sample of M's averaging window, because phase must not be averaged and the two arms' frames must still line up.

L (learned) used 24 sinc band-pass filters with two learnable cutoffs each, kernel length 0.5 s, followed by squaring, average pooling to 25 Hz, and a logarithm. It pools power before the logarithm, unlike M, which averages log power. Filter shape and normalization also differ, so learnability is not isolated. Two initializations were run, one on M's frequency grid and one with uniform random cutoffs in 1 to 45 Hz.

M-sub is M's log power decimated by subsampling rather than averaging, at M-phi's offset so all three arms' frames align, at the same 25 Hz output rate. **M-phi-scram** is M-phi's coefficients with a content-derived rotation per trial and per frequency, shared over time and channels. It preserves pre-z-scoring coefficient magnitudes and within-frequency phase differences, while altering cross-frequency relationships and alignment across trials. The rotation is content-addressed, so it is a fixed property of an epoch rather than a stochastic augmentation, and identical across seeds.

The angle is computed from the epoch's own samples, not drawn from a generator:

```
probe = flatten(epoch)[::997]                    # prime stride
key   = sum(probe * 1e6) + 20260822              # study-wide salt
theta = frac(sin(key * 12.9898 + j * 78.233) * 43758.5453) * 2*pi
```

for filter index j, with the probe cast to double precision. The prime stride avoids aligning with the channel or sample periodicity of the epoch layout, and the mixing constants are a standard fractional-sine hash: distinct per trial and frequency, stable for identical input, and free of any generator state that could differ across processes. A fresh per-forward draw would instead make the arm a stochastic augmentation and confound the contrast against M-phi, which has no such noise, with a regularization effect.

Label independence is an assumption, not a guarantee of this deterministic function of the signal. The registered P2-A diagnostic (E15) used 600 real P300 epochs from three people, with 481 standard and 119 target trials, matching the archived preprocessing configuration and 29-channel layout. For each of 24 filters it compared the two classes' circular mean resultant lengths, using their absolute difference against 10,000 pooled-epoch label permutations with class sizes fixed. Bonferroni correction used family alpha 0.01; the smallest corrected p-value was 0.113, so this test detected no class difference. The status is inconclusive: equal resultant lengths do not establish equal angle distributions, the local sample is not a headline population, and the check did not read the production runs' angle tensors. It neither proves label independence nor changes C12's registered contrast-ordering result. **D2** applies a fixed random permutation to the frequency axis.

B.4 S2.4 NORMALIZATION SCOPE, AND THE CONFOUND IT INTRODUCES

Normalization differs by arm, by design, and it is a limitation rather than a detail. Raw used per-recording robust scaling, the convention from the sleep staging literature; the spectral arms used per-frequency z-scoring fit on the training population, the convention from the spectral-parameterization literature; the learned arms used batch normalization over band channels, because fixed per-band statistics are ill-defined while cutoffs move during training.

The consequence runs in a direction worth naming. A per-subject gain becomes an additive offset in log power, which pooled z-scoring does not remove, while per-recording scaling divides it out.

The arm contrast would therefore bundle representation with normalization scope, and a reader would be entitled to attribute some of the raw arm's advantage to its normalization. S5.5 reports the control that separates them, run in both directions; both matched comparisons retain a substantial deficit.

B.5 S2.5 Corpora, labels, sampling and retained counts

The recorded source identifiers are NEMAR (Delorme et al., 2022) on007056 (PURSUE VisualOddball), on007096 (PURSUE FacePerception) (Couperus et al., 2026b;a), and nm000225 (PhysioNet/CinC Challenge 2018 sleep annotations). The ERP event-map records and the sleep metadata were inspected directly. The archived experiment records do not recover an exact upstream release tag or source-metadata commit for these acquisitions. The PURSUE accession citations identify the sources, not verified historical acquisition snapshots. The accessible PhysioNet release 1.0.0 identifies the original challenge source (Ghassemi et al., 2018); it is not evidence that this precise release was the NEMAR snapshot used by our runs.

For P300, repeated consecutive digits are targets and other mapped digit pairs are standards; response, boundary and unmapped event codes are excluded. For N170, intact faces (codes 1–40) and cars (41–80) are retained; scrambled images, button/response events and boundaries are excluded. No extra class balancing or artifact rejection is applied to these ERP epochs.

Sleep uses Wake, N1, N2, N3 and REM labels, excluding concurrent arousal and respiratory markers. The dated construction record identifies 903 eligible people with all five stages among the 994 annotated challenge records, then a random cohort of 320 at seed 20260822. Nonoverlapping one-second windows are tiled within consolidated stage runs, without straddling stage boundaries. The six EEG signals are F3-M2, F4-M1, C3-M2, C4-M1, O1-M2 and O2-M1; other polysomnographic channels are excluded before the common preprocessing. These are one-second sleep-segment labels, not a conventional 30-second staging benchmark.

Sleep epochs are capped at 300 per person using equal available-class stratification without replacement. Scarce classes contribute their available windows; remaining quota is distributed among richer classes using integer division, which can leave a small unfilled remainder. The preserved entrypoint and batch script default to epoch-sampling seed 1 per person. The historical launched override is not recovered, so the configured default is not asserted as verified invocation provenance. This differs from the cohort-selection seed above.

Each paradigm has 40 fixed evaluation people selected before the primary grid, excluded from its training and validation draws. The following counts were read from every retained primary-sweep label bundle; they are identical across cells within each paradigm. The metadata audit verifies the archive checksums before reading labels (E01; `cohort-metadata.json`).

Evaluation task	People	Retained epochs	Counts by class	Epochs/person
P300	40	8277	standard 6621; target 1656	165–210
N170	40	6517	car 3146; face 3371	50–166
Sleep	40	11998	Wake/N1/N2/N3 each 2407; REM 2370	298–300

Complete training-cohort class/trial totals and upstream release manifests are not recoverable from the local archive. These fixed evaluation counts must not be substituted for training totals. Training draws are nested within each of seeds 1, 2 and 3 across the subject-count grid; variation across those fits therefore includes training subject selection as well as optimization randomness. The conditional evaluation-person intervals do not estimate that variation.

B.6 S2.6 Optimization and the learning-rate grid

Identical for every arm: AdamW with weight decay 0.05, a cosine schedule, batch size 64, 30 epochs, and best-epoch selection on a validation split formed as 85/15 by subject within the n-person draw, not by trials. Thus n includes internal validation people. Every archived primary cell records the following allocations:

n	Training people	Validation people
20	17	3
40	34	6
80	68	12
160	136	24
240	204	36

Pooled normalization statistics use the full n-person draw, including internal validation people, as specified by the normalization amendment; evaluation people never fit those pooled statistics. The learning-rate grid was {3e-4, 1e-3}, identical for every arm and selected on validation.

Two points cannot bracket an interior optimum, so no claim here says the learning rate was optimized. It was selected from two candidates, and the property protecting the comparison is that the grid was the same for all arms. Measured over arm R on the visual oddball paradigm, 15 cells: selections split 6 at 3e-4 and 9 at 1e-3, four of the five subject counts select both rates rather than pinning at one end, and the validation gap between the two rates has median 0.0066 balanced accuracy, minimum 0.0010 and maximum 0.0268. These describe only the audited R/P300 cells; they neither establish that other arms were well bracketed nor replace a wider search.

B.7 S2.7 The frequency-structured embedding

The convolutional arms consume bit-identical tokens to their linear partners, from the same front-end class with the same settings. Only the embedding slot differs. Tokens are folded back into a per-channel time-frequency plane, convolved, pooled into frequency bands, and unfolded:

```
(B, n_ch*n_frames, C_in*n_freq)
  -> (B*n_ch, C_in, n_freq, n_frames)
  -> Conv2d(C_in, 8, kernel=(5, 3), padding=same)
  -> GELU -> average-pool frequency to 4 bands
  -> (B*n_ch, 8, 4, n_frames) -> flatten (8, 4)
  -> Linear(32, d_model)
  -> LayerNorm (no learnable affine)
```

`C_in` is 1 for the log-power arms and 2 for M-phi, whose real and imaginary halves become the convolution's two input channels. Token count is preserved exactly, so the channel and time positional embeddings, the transformer body, the pooling and the head are untouched.

Pooling to **four bands rather than one** is the difference between band power and broadband power, and pooling to one is the defect described in S4.4. The LayerNorm carries no learnable parameters and exists only to match the output scale that the additive positional embeddings operate at.

B.8 S2.8 The self-supervised arm

Masked-reconstruction pretraining ran on pooled continuous unlabeled recordings, restricted to subjects outside every evaluation population: 230 distinct subjects, 679 subject-by-paradigm records, 111,566 epochs, after a cross-paradigm leakage fix.

Pretrained weights transferred to the same evoked paradigms at two subject counts under two regimes, linear probing and full fine-tuning, because the literature documents ranking reversals between them (BrainBERT, Section 2.2). The observable is the pretraining delta per front end, pretrained minus from-scratch, and then the contrast between those deltas. In the archived block (E05) **the**

from-scratch side is not recomputed; it is the same arm, paradigm, subject count and seed from the main sweep.

The reported native-target numbers come from a re-run of the whole fine-tuning block (E12): 165 executions on one four-GPU node, in one allocation and at one code revision, under deterministic algorithms. The from-scratch side is recomputed inside the block, and seeds 4 to 6 are added on both sides, giving 72 fine-tune and 72 from-scratch cells over six seeds beside the 12 budget-matched cells. All six seeds fine-tune from the same pretrained model per arm, so the interval samples training draws and optimization, not pretraining runs, and seeds 4 to 6 redraw people from the same pool rather than reaching new ones. Six same-device repeats of the 80-subject, second-seed cells agree exactly on every per-epoch value and on evaluation balanced accuracy. Three cross-device repeats measure the device: one differs in the sixth decimal place of its training curves with identical evaluation accuracy, so each seed composite carries a device component that is not separated. The linear-probing regime was not re-run.

Each arm reconstructs its own native tokens, so the arms do not share a prediction target, and a registered check measured that the native tasks differ in kind (C4-E4P; E13). We expected the complex-coefficient target to be the easy one, since the coefficients are a band-limited, oversampled encoding of the same 200 samples; the measurement shows the opposite, and corrects the mechanism we had stated. The check used 4,096 epochs from the pretraining pool, which every model saw under other masks and which, if anything, flatters the models, and one fixed mask per arm. Interpolability, one minus the error of linear interpolation along time from the nearest unmasked frames (samples, for raw) over the error of predicting zero, is 0.77 for a masked log-power token and 0.22 for a masked raw sample. For a masked complex coefficient it is -0.71: interpolation does worse than predicting zero, because the coefficient's phase advances too far across a masked gap for a straight line to follow. The registered comparison of each pretrained model against that interpolator read PARTIAL: the model's error over the interpolator's is 0.52 for the phase-preserving arm and 0.13 for raw, where the premise needed at least 0.8 and at most 0.5. In an exploratory reading made after the record was seen, the phase-preserving model recovers 0.11 of the masked coefficients' energy, against 0.90 for raw and 0.94 for log power, and in every band from 9.8 Hz up it recovers at most 1.6 percent. The complex target was therefore the harder task, not the easier one. Equal nominal masked error does not mean equal task difficulty, and a larger pretraining gain for the raw arm cannot be attributed to phase. This bounds the contrast's interpretation whatever its interval.

Learning-rate grids for this arm were widened several times and every widening is recorded with its reason. The final grids were 10 points for probing and 7 for fine-tuning. Selections were audited for endpoint values before any number was read. **The linear-probing regime is reported as inconclusive**: after widening, its selected rates still spanned four orders of magnitude across twelve cells and one cell remained pinned at an endpoint, which indicates an objective whose scatter exceeds its range rather than a rate that needs finding.

Because the pretrained arms searched a wider grid than the from-scratch cells they are differenced against, a control restricted both sides to the identical two-point grid.

C4-E4 uses a common waveform target (E16). It masks 40-sample (200 ms) blocks of a channel before applying any front end, selecting 72 of 145 blocks per 29-channel epoch with a fresh mask at each step. Masked samples are zeroed in the input, and the corresponding tokens receive the learned mask embedding. Every arm predicts the same per-recording robust-scaled raw samples, with masked-sample mean squared error as the loss. The reconstruction head emits 20 samples per raw token or 8 per spectral token. Zeroing before the wavelet transform prevents unmasked spectral tokens from containing masked target samples, but also removes context from wavelets overlapping masked blocks. This context handicap favors raw; the intervention changes masking domain and geometry as well as prediction target.

The spectral arms' frequency-z-score statistics were fit on 4,096 unmasked pool epochs, sampled by pretraining seed, then applied to zero-filled inputs. R's input is already in the common target's per-recording robust-scaled units. The spectral arms instead use pooled frequency normalization and must infer the per-recording gain from one epoch. At fine-tuning, the spectral arms move from zero-filled pretraining inputs to clean inputs, while R's unmasked tokens do not change.

Pretraining retains the evaluation-excluded pool of 679 records, 100,000 steps, batch size 128, mask ratio 0.5, size M, AdamW at 3e-4 with weight decay 0.05, and cosine scheduling. Six seeds per arm

produce 18 checkpoints. Each is fine-tuned at the matching draw seed on P300 and N170, at n = 20 and 80; scratch comparators are rerun inside the same block. The 72 fine-tune cells, 72 scratch cells and 18 pretraining runs give 162 executions, all under deterministic algorithms at one deployed revision. Seeds 1–2, 3–4 and 5–6 occupy three separate nodes; every contrast within a seed stays within its node. The seven-point fine-tune and two-point scratch grids remain asymmetric. The amended manipulation check requires every pretrained model to beat both a zero predictor and a linear interpolator over the final 10,000 training steps; a failed run permits neither a replacement seed nor a primary reading. S5.3 reports that check, the transfer reading and the registered sensitivities.

B.9 S2.9 Probes of published checkpoints

E07 identifies the LaBraM-Base and REVE-Base regression reports. Both use on007056 VisualOddball and the same recorded evaluation-population manifest, SHA-256:

`4d3d96508c60038ef503e5ce287939409db15dbcc2520008c429150e9a4e2427`.

The 40 people are split by person into 24 training, 8 validation and 8 test people at seed 20260819, with 768, 256 and 256 selected epochs, respectively. The preserved runner samples epochs uniformly without replacement, with a default cap of 32 per person; its historical cap override is not recorded. Each checkpoint uses its own preserved preprocessing pipeline rather than the scratch-arm tensor: `labram_adr0004_own_pipeline` with configuration `labram_patch200_sfreq200.0`, or `reve_adr0004_own_pipeline` with `reve_patch200_sfreq200.0`. Input-provenance hashes are a4627bd1ff9e and 62cd1ac0cf85, respectively. The spectral target is computed from the same own-pipeline epoch supplied to that checkpoint, not a separately preprocessed scratch-arm epoch.

The pooled embeddings have widths 200 for LaBraM and 512 for REVE. A ridge readout predicts Morlet log power averaged over channels and frames, against a matched random-initialization encoder using the same architecture, people, epochs and seed. The six candidate penalties are $\{10^{-2}, 10^{0}, 10^{2}, 10^{4}, 10^{6}, 10^{8}\}$. LaBraM selects 10^{4} for the trained checkpoint and 10^{2} for its null; REVE selects 10^{2} for both. Selection uses validation people; test people determine the reported trained-minus-null result.

LaBraM checkpoint SHA-256:

`7c50583826afac76c4ab18f43d958df40496c8229accc09ed6a227c9bb57c37c`.

REVE checkpoint SHA-256:

`8ecc650619598748286c2457f81f5c6bd12e8bb59db44f7b02af1955c44de8fe`.

The checksums identify the checkpoint files actually probed. The archive does not establish their complete upstream pretraining-data versions or upstream training-code commits. Likewise, the input-population hash identifies our selected people, not a recoverable dataset release.

B.10 S2.10 Perturbation and spectral-response probes (C7, C8)

Two further probes of the same published checkpoints are registered in the ledger and are reported here so that every ledger row has visible evidence. Neither supports a claim in the main text.

C7 asked whether published raw-input models' input stages are band-selective, using a perturbation probe: per-band embedding cosine shift under band-notch and band-pass-only perturbations, against a matched random-initialization null. The two checkpoints disagree. REVE's band-pass-only shifts exceed its null by a large factor, peaking in theta and alpha, which is real selectivity. LaBraM's band-notch shifts are *smaller* than its null in delta, theta and alpha, that is, the trained model is less disturbed by removing a band than the untrained one is. The notch and band-pass directions also disagree for LaBraM. C7's status is `inconclusive` and it is not quotable as a general statement about raw-input models.

C8 originally claimed both input stages peak at high frequency. That phrasing failed: REVE's peak sits exactly at the analyzable band edge, so it is a lower bound rather than a tuning frequency, which is the same defect the results gate checks for elsewhere. The claim was rephrased before any number reached a draft, to the weaker statement that pretraining moves both checkpoints' aggregate

input-stage response mass higher in frequency than the untrained architecture places it, with the band-edge case reported as a lower bound and never as a frequency. Because the rephrasing followed the numbers, C8's status is `exploratory` and it supports no confirmatory claim.

Taken with C5 and C6, these probes measure heterogeneity between published checkpoints rather than a battery supporting this study's thesis: of the four, only C6 points the thesis-favorable way and C5 points partly against it. Both probe reports are retained at probe code revision `cddebdf0e1a4a3a999a99b2260b001c23c5c40b0`.

C S3. Statistical specification and per-paradigm tables

C.1 S3.1 Estimand and dependence correction

For each fitted model, balanced accuracy is the unweighted mean of class recalls computed from pooled evaluation trials. The arm contrast is calculated within a seed, then averaged equally over the three fitted seeds. The ERP composite equally weights P300 and N170. It is not a mean of subject-specific balanced accuracies.

The revised intervals use evaluation people as clusters and keep their identity paired across every contrast. P300 and N170 each contain 40 evaluation people: 18 appear in both tasks, 22 only in P300, and 22 only in N170. Each bootstrap replicate samples within those three membership strata with replacement, retaining the observed count in each stratum. A shared person's multiplicity is reused in both tasks. Sleep has a separate namespace and samples its 40 evaluation people. Every selected occurrence includes that person's evaluation trials.

One deterministic resampling plan is reused across fitted seeds, arms, subject-count levels, and the SSL/scratch contrasts. Each replicate recomputes pooled-trial balanced accuracy, forms the paired contrasts, then takes the same equal-weight averages as the point estimator. The reported interval uses the 2.5th and 97.5th percentiles, with linear interpolation, of 1,000 replicates at master seed 20260819. Paired changes across training scales use the same person multiplicities at both levels; subtracting separately computed interval endpoints would be invalid.

C.2 S3.2 What the intervals cover

These are conditional evaluation-person intervals. They condition on the fitted models, selected training draws and observed cohort membership counts. They do not estimate variation from retraining, hyperparameter selection, new training-subject draws, or a different mixture of task memberships. More trials do not create more independent people, but can still improve within-person measurement; the previous assertion that additional trials could never narrow uncertainty was too strong.

The seed spread alongside each interval is descriptive variation across the three available fits. It is not a confidence interval or an estimate of total uncertainty. We do not add a separately estimated seed variance to the corrected bootstrap: the previously tabulated augmented-variance sensitivities used the superseded dependence scheme and remain historical outputs.

The original analysis sampled the same evaluation people independently across fitted seeds and did not preserve shared membership across the two ERP tasks. The post-review amendment corrects those dependencies without changing the pooled-trial point estimator. It was specified after the original results were known and independently reviewed before manuscript migration; it is not a blind pre-registration of a new experiment. The corrected bundle preserves all audited point occurrences while changing several interval exclusions. In particular, the sleep gap at n = 160 no longer excludes zero, and the small averaging contrast at n = 240 now excludes zero negatively. Neither correction changes the large ERP deficit or the dominant rotation contrast.

C.3 S3.3 Readings, multiplicity and numerical resolution

The original registered readings specify their subject-count levels and signs (S1). Intervals are pointwise, not simultaneous bands over all subject counts, arms, or claims. The report also retains secondary per-paradigm Holm adjustments within each subject-count level; these do not provide

familywise coverage of the complete study. We report effect sizes and intervals here, without treating the finite bootstrap tail proportions as exact p-values.

The committed resolution audit repeats the corrected C17 reading with 1,000, 10,000 and 50,000 shared-person resamples and returns the same supported reading. That check concerns numerical resolution under the conditional estimator, not training-seed uncertainty. C4-E3's native-target interval is a different estimator, a t interval over six training-draw seeds that carries a device component (S5.3). All draws use one pretrained model per arm, so it does not include pretraining variation. C4-E4 uses the same per-seed composite, averaging the two ERP paradigms and two subject counts, but each draw uses its own pretrained model per arm. Its 95 percent t interval has five degrees of freedom and multiplier 2.571; it carries pretraining-seed, node and device components. The three node groups hold two seeds each, so node and seed-pair effects are not separable. The registered between/within-pair diagnostic and a descriptive interval over the three pair means are reported in S5.3; neither replaces the primary reading. E3 and E4 therefore have different variance structures and are compared only descriptively.

C13-T uses the conditional subject-cluster intervals above. Its SURVIVES reading at n = 20 and 40 requires both matched gaps to keep the within-block unmatched gap's direction and at least half its point magnitude; intervals do not enter that decision (S5.6). Per-seed spreads remain descriptive. P2-A is a separate permutation diagnostic of classwise resultant lengths, with Bonferroni correction over 24 filters, not an omnibus test of angle-label independence (S2.3).

C.4 S3.4 Corrected effect tables

The following tables are formatted directly from the independently reviewed subject-cluster report bundle. The versioned source paths and checksums are recorded with the generated tables and in the revision map. All earlier reports are retained as historical evidence.

Intervals condition on the fitted models, selected training draws and observed cohort membership counts. Seed spread is the reported maximum-minus-minimum of three fitted-seed points, not an interval.

Task	Contrast	Claim	n	Point	Conditional 95% CI	Seed spread
ERP	M minus R	C1	20	-0.0925	[-0.1098, -0.0750]	0.0198
ERP	M minus R	C1	40	-0.0909	[-0.1075, -0.0737]	0.0075
ERP	M minus R	C1	80	-0.1126	[-0.1316, -0.0948]	0.0229
ERP	M minus R	C1	160	-0.1096	[-0.1258, -0.0946]	0.0058
ERP	M minus R	C1	240	-0.1122	[-0.1273, -0.0966]	0.0131
P300	M minus R	C1	20	-0.0462	[-0.0663, -0.0281]	0.0278
P300	M minus R	C1	40	-0.0645	[-0.0834, -0.0459]	0.0223
P300	M minus R	C1	80	-0.0835	[-0.1049, -0.0641]	0.0409
P300	M minus R	C1	160	-0.0815	[-0.0997, -0.0644]	0.0195
P300	M minus R	C1	240	-0.0821	[-0.0978, -0.0663]	0.0283
N170	M minus R	C1	20	-0.1388	[-0.1665, -0.1136]	0.0477
N170	M minus R	C1	40	-0.1173	[-0.1388, -0.0946]	0.0162
N170	M minus R	C1	80	-0.1417	[-0.1656, -0.1190]	0.0108
N170	M minus R	C1	160	-0.1377	[-0.1609, -0.1164]	0.0080
N170	M minus R	C1	240	-0.1423	[-0.1639, -0.1206]	0.0157
ERP	M-phi minus M	C11	20	+0.0553	[+0.0444, +0.0662]	0.0968
ERP	M-phi minus M	C11	40	+0.0795	[+0.0657, +0.0921]	0.0449
ERP	M-phi minus M	C11	80	+0.0983	[+0.0820, +0.1139]	0.0318
ERP	M-phi minus M	C11	160	+0.0895	[+0.0744, +0.1042]	0.0134
ERP	M-phi minus M	C11	240	+0.0942	[+0.0805, +0.1067]	0.0262
P300	M-phi minus M	C11	20	+0.0245	[+0.0124, +0.0364]	0.0179
P300	M-phi minus M	C11	40	+0.0401	[+0.0293, +0.0514]	0.0635
P300	M-phi minus M	C11	80	+0.0623	[+0.0482, +0.0757]	0.0558
P300	M-phi minus M	C11	160	+0.0567	[+0.0437, +0.0702]	0.0260
P300	M-phi minus M	C11	240	+0.0569	[+0.0431, +0.0696]	0.0478
N170	M-phi minus M	C11	20	+0.0862	[+0.0680, +0.1056]	0.1833
N170	M-phi minus M	C11	40	+0.1188	[+0.0949, +0.1411]	0.0264
N170	M-phi minus M	C11	80	+0.1344	[+0.1086, +0.1605]	0.0135

Task	Contrast	Claim	n	Point	Conditional 95% CI	Seed spread
N170	M-phi minus M	C11	160	+0.1224	[+0.0984, +0.1468]	0.0260
N170	M-phi minus M	C11	240	+0.1315	[+0.1085, +0.1543]	0.0178
ERP	M-phi minus rotated	C12	20	+0.1199	[+0.1052, +0.1344]	0.1039
ERP	M-phi minus rotated	C12	40	+0.1521	[+0.1354, +0.1672]	0.0091
ERP	M-phi minus rotated	C12	80	+0.1682	[+0.1502, +0.1861]	0.0233
ERP	M-phi minus rotated	C12	160	+0.1820	[+0.1646, +0.2000]	0.0410
ERP	M-phi minus rotated	C12	240	+0.1955	[+0.1772, +0.2121]	0.0107
ERP	rotated minus M-sub	C12	20	-0.0620	[-0.0721, -0.0521]	0.0215
ERP	rotated minus M-sub	C12	40	-0.0725	[-0.0825, -0.0625]	0.0464
ERP	rotated minus M-sub	C12	80	-0.0706	[-0.0810, -0.0612]	0.0218
ERP	rotated minus M-sub	C12	160	-0.0864	[-0.0968, -0.0763]	0.0422
ERP	rotated minus M-sub	C12	240	-0.0974	[-0.1107, -0.0853]	0.0060
ERP	M-sub minus M	C12	20	-0.0025	[-0.0061, +0.0012]	0.0172
ERP	M-sub minus M	C12	40	-0.0001	[-0.0043, +0.0044]	0.0105
ERP	M-sub minus M	C12	80	+0.0007	[-0.0040, +0.0058]	0.0137
ERP	M-sub minus M	C12	160	-0.0061	[-0.0103, -0.0017]	0.0043
ERP	M-sub minus M	C12	240	-0.0039	[-0.0075, -0.0004]	0.0201
Sleep	M minus R	C13	20	+0.0220	[+0.0093, +0.0342]	0.0364
Sleep	M minus R	C13	40	+0.0294	[+0.0128, +0.0450]	0.0097
Sleep	M minus R	C13	80	+0.0136	[+0.0019, +0.0246]	0.0095
Sleep	M minus R	C13	160	-0.0068	[-0.0173, +0.0023]	0.0048
Sleep	M minus R	C13	240	-0.0060	[-0.0175, +0.0050]	0.0164
Sleep	M-phi minus M	C13	20	-0.0364	[-0.0459, -0.0267]	0.0104
Sleep	M-phi minus M	C13	40	-0.0324	[-0.0427, -0.0213]	0.0140
Sleep	M-phi minus M	C13	80	-0.0243	[-0.0304, -0.0181]	0.0105
Sleep	M-phi minus M	C13	160	-0.0057	[-0.0124, +0.0000]	0.0071
Sleep	M-phi minus M	C13	240	-0.0066	[-0.0130, -0.0005]	0.0101

D S4. Negative, void, and superseded results

D.1 S4.1 Frequency permutation and the exploratory learned-filter observation

D2 applies a fixed frequency permutation to M's spectral column. For a learned linear projection, the permutation can be absorbed into its weights: the permuted embedding computes (WP)x instead of Wx. With exchangeable initialization across input coordinates, these model families are equivalent in distribution, though corresponding fitted seeds need not follow identical trajectories. The control therefore does not isolate the value of frequency order for this embedding.

Across the ten ERP cells, the M-minus-D2 difference alternates in sign. Its maximum absolute difference is about 0.0120 balanced accuracy, below M's median three-seed spread of 0.0161. This is consistent with training variability; it does not validate the registered denoising interpretation. C17 instead tests permutation under a frequency-structured convolution (S6).

L-random is a trainable sinc filterbank initialized at random cutoffs. It differs from M in filter construction and normalization as well as trainability. Its small observed cutoff movement and competitive performance are exploratory observations, not a fixed filterbank-identity control or a clean test of learnability alone.

D.2 S4.2 Unsupported original readings

C1 required a positive M-minus-R ERP composite at n = 20, 40 and 80. The corrected intervals are negative throughout (S3). An earlier single-seed screen had already shown a raw-favored sign, and the reading was not changed. That screen's run directory was not retained; it is historical context, not independent confirmatory evidence.

C3's learned-filter convergence prediction failed. L-random minus L-band is positive on N170 at every subject count, from +0.024 to +0.053. The original rule also failed to distinguish convergence from a layer that barely changed.

C4's original fine-tune contrasts were -0.0039 for M and -0.0248 for M-phi. The M-phi contrast was negative in all four cells and -0.0409 under the identical-grid control, and the normal two-standard-error exclusions were withdrawn. These archived values are superseded as the reported numbers by the deterministic re-run (S5.3; E12), under which the M-phi contrast keeps its sign with a six-seed interval that carries a device component, and the M contrast's interval includes zero, so no direction is reported for M. The archived reruns' drift from their nominal twins is kept as a record (E06); the re-run removes it as a property of the reported numbers without explaining why it concentrated in M-phi. The arms' native reconstruction targets are still not a common objective, so neither block establishes a transfer effect of phase (S2.8, S5.3; E05, E06). The separate common-target E4 experiment is inconclusive as E4-NULL and does not identify the cause of the native-target contrast (E16).

C10's two stacks agree in the raw-favored sign at every subject count, but its original mutual-interval-containment reading failed at one level. That containment analysis has not been migrated to the corrected estimator and is retained only as history; S5.2 reports the between-stack differences that replace it (E09).

D.3 S4.3 Withdrawn spectral-probe result

C9 reported R-squared values -2.76 and -4.19 from ridge fits without an intercept. Those values were withdrawn and replaced by the corrected trained-minus-random checkpoint comparisons in S2.9 (E07). The original values remain in the corrections table and ledger, but do not enter the active evidence for spectral accessibility.

D.4 S4.4 Failed convolution implementation and correction

An early complex-convolution validation run on P300 at n = 20, seed 1, returned exactly chance at epoch zero. The run directory was overwritten during correction, so the result survives only in the dated design amendment, not as a reproducible archived run. It is not part of the confirmatory block.

The implementation pooled over all frequency bins, removing band identity, and delivered activations much smaller than the additive positional embeddings. The retained scale diagnostic reproduces a standard deviation of 0.0854 after single-band pooling versus 0.5886 for the linear embedding. These are diagnostic outputs from the reconstructed implementation, not recovered measurements from the missing run.

The correction pools into four bands and applies non-affine layer normalization, bringing the diagnostic output standard deviation to 0.9998. The same correction was applied to all convolutional arms before their confirmatory runs. The retained validation run reaches 0.6455 validation and 0.6242 evaluation balanced accuracy at epoch 23 (E10). The corrected model refuses a pooling configuration with fewer than two bands.

This establishes that the corrected path trains on the validation cell. It does not prove that the correction is neutral between all architectures or remove the need for the subsequent matched controls.

E S5. Robustness and controls

E.1 S5.1 Objection: the regime difference is a channel-count artifact

The evoked corpora carry 29 electrodes and the sleep corpus 6, and the front-end verdict inverts between them. A spectral summary discards less, relatively, when there is less to discard, so the inversion could track the montage rather than the regime.

The control was pre-registered before it ran: one evoked paradigm re-run end-to-end on six electrodes, everything else identical.

n	G(n), six electrodes	G(n), full montage
20	-0.0659	-0.0462
40	-0.0464	-0.0645
80	-0.0603	-0.0835

n	G(n), six electrodes	G(n), full montage
160	-0.0663	-0.0815
240	-0.0509	-0.0821

The evoked gap survives at every subject count, running from -0.0464 to -0.0663 against -0.0462 to -0.0835 on the full montage. Reducing the evoked paradigm to the sleep montage's channel count does not reproduce the sleep result: the sleep curve is positive at small subject counts and crosses zero, while this one stays negative throughout.

What it does not settle: these six electrodes are a subset of an evoked montage, not the sleep montage, and the paradigms still differ in every other respect. The control rules out channel count as a sufficient explanation without isolating what in the regime difference does the work (S6).

E.2 S5.2 Objection: the result is a property of your compute stack

An unattended driver upgrade once split one of our sweeps across two software stacks, and nothing in the run records would have revealed it. That made stack drift detectable but did not establish that the stack is irrelevant.

The comparison covers one paradigm over 54 cells completed on both stacks: an Ada-generation consumer GPU on one CUDA build against a Volta datacenter GPU on an older one, differing in both hardware generation and toolchain. The sign agrees at all five subject counts and both stacks are raw-favored. The historical interval-containment reading is not part of the corrected subject-cluster bundle and supplies no current interval claim (S4.2).

The descriptive quantity is the between-stack difference. Over the 54 shared cells the per-cell absolute difference in balanced accuracy has median 0.0065, 95th percentile 0.0218 and maximum 0.0317. That is how far a headline result of this kind moves across deliberately different stacks.

Provenance note. The comparison that first produced these numbers was run without writing its output to a file, so the table in our records was for some weeks a hand transcription with no artifact behind it. It has since been regenerated from both sweeps and every value reproduces exactly, including the single containment failure. The point is not that the transcription was faithful but that we could not have known it was until we checked (S1.3).

E.3 S5.3 Pretraining contrasts: search budget and prediction target

The pretrained arms searched a wider learning-rate grid than the from-scratch cells they are differenced against, because their grid had to be widened several times before it bracketed anything. A wider search on one side of a difference can manufacture the difference.

The control restricts both sides to the identical two-point grid and recomputes the contrast. Both blocks are shown, the archived one (E05) and the deterministic re-run (E12), with every term of a column from its own block:

contrast	archived, as run	archived, budget-matched	re-run, three seeds	re-run, budget-matched
d[M] - d[R]	-0.0039	-0.0103	-0.0085	-0.0156
d[M_phi] - d[R]	-0.0248	-0.0409	-0.0192	-0.0462

The M-phi contrast is negative in all four cells of every column. Budget matching makes it more negative in both blocks, so search-grid asymmetry is not the explanation. The asymmetry itself is reproduced by the re-run, not repaired: a two-point grid makes every from-scratch selection an edge. The archived composites also depend on which of the two archived runs of a block is read, so an archived value is quoted with its block (E05).

Six seeds give the reported interval. The re-run adds seeds 4 to 6 on both sides of the contrast. Over six seeds the M-phi composite is -0.021068, with a 95 percent t interval of [-0.0396, -0.0025] (t =

2.571, five degrees of freedom), which excludes zero. It is a seed interval: its estimand, fixed before the re-run, is training-draw and optimization variance under exact same-device repeats, and it carries the unseparated device component described in S2.8. The M composite is -0.0077 with an interval of [-0.0200, +0.0046], so no direction is reported for M. Two caveats cut against the reading. The exclusion is narrow, and seed 3's composite is +0.0058; an exploratory count made after the interval was read finds that seed 3 also holds both M-phi fine-tune cells whose selected rate sits at the top edge of the seven-point grid, which licenses neither dropping the seed nor re-selecting the cells. And six seeds fine-tuned from one pretrained model per arm say nothing about variation across pretraining runs.

The earlier normal two-standard-error exclusions over four archived cell means remain withdrawn (S4). The archived reruns (E06) drifted from their nominal twins at the scale of the contrast; that record stays in the claim gate as a failing check, and the re-run replaces those numbers rather than explaining the drift.

Separately, each arm's pretraining reconstructs its own native tokens, which is not a common objective across arms (S2.8), so even an exactly reproducible contrast with an interval does not establish whether pretraining favors phase.

The common-target experiment reads E4-NULL (C4-E4; E16). All 162 registered executions completed at one clean deployed revision. Every one of the 18 pretrained models beat both the zero predictor and registered linear interpolator over the amended final 10,000 training steps, with 201 logged points per run in that window. No seed was replaced, so the manipulation check permits the reading.

Seed	1	2	3	4	5	6
d[M-phi] minus d[R]	-0.01680	-0.00355	+0.03721	+0.03198	-0.00569	+0.01023

The mean is +0.008896, with a six-seed 95 percent t interval of [-0.013926, +0.031719], carrying pretraining-seed, node and device components. The registered primary reading is E4-NULL: no direction is reported, and an interval containing zero does not establish equality or equivalence. The registered secondary contrast, d[M] minus d[R], is -0.014140 with a six-seed 95 percent t interval of [-0.024298, -0.003982], carrying the same components. It reads E4-NEGATIVE and moves no primary letter.

The three node/seed-pair means are -0.010176, +0.034593 and +0.002272. The registered between/within-pair F is 14.0495, above the F(2, 3) threshold 9.552, so a node component is reported. The descriptive t interval over three pair means is [-0.048512, +0.066305] at two degrees of freedom, also including zero; this diagnostic does not replace the registered six-seed interval or change its letter.

E3's native-target mean is -0.021068 and E4's is +0.008896, a descriptive difference of +0.029964. E4's per-arm gains are +0.032748 for M-phi, +0.009711 for M and +0.023851 for R; E3's are +0.000201, +0.013576 and +0.021269. Different objectives, masking geometries and variance structures prevent reading these differences as a test of why E3 differs.

One scratch cell, M-phi on N170 at n = 20 and seed 3, is exactly chance. The registered point sensitivity drops its matched N170/n = 20 slice: the composite changes from +0.008896 to +0.000184. This is a sensitivity point, not a second interval, and the primary retains the cell. Two of 72 fine-tune selections lie at grid edges, below the registered concentration rule. The wavelet-context handicap and seven-point versus two-point search-budget asymmetry remain. Spectral normalization statistics were fit on 4,096 unmasked pool epochs, then applied to zero-filled inputs. R's input is already in the common target's per-recording robust-scaled units. The spectral inputs instead use pooled frequency normalization and must infer the per-recording gain from one epoch; their fine-tuning inputs also change from zero-filled to clean. C4 stays not supported on its native reading; E4 is inconclusive and identifies neither a phase advantage nor the cause of E3.

E.4 S5.4 One degenerate cell, and the direction it biases

One cell of the sweep returned exactly chance: the phase-preserving arm on face perception at 20 subjects, third seed, a collapsed fit predicting a single class, where the same cell's other two seeds reached 0.7131 and 0.6963. It is the only cell of the headline sweep at chance **relative to its own arm**, and no pretrained cell is degenerate. The conv block adds two more in the same arm family, paradigm and subject-count range; S6.4 reports all three together because the pattern across them is the interpretable part.

Two other cells also sit at exactly chance and are not degenerate. Both belong to the phase-scrambled arm on face perception, whose entire series runs between 0.497 and 0.561 across every seed and subject count. That arm is a deliberately perturbed representation, so its poor performance is not evidence of a collapsed fit relative to successful sibling seeds.

We report the degenerate cell rather than dropping it, and state which way it pushes. The self-supervised observable is a difference, pretrained minus from-scratch, so a from-scratch score that is too low inflates that arm's apparent benefit and pulls the contrast toward zero. The cell therefore works **against** the direction C4 reports. Excluding that run moves the archived contrast (E05) to -0.0295, negative in all four cells; the sensitivity's previously reported two-standard-error exclusion was withdrawn, since its corrected t interval over four cell means includes zero. No exclusion is claimed. The sensitivity is regenerated rather than quoted, and it drops the single degenerate run, which is a different operation from dropping that seed everywhere.

The phase-scrambled arm is the comparator in the mechanism factorial, so we checked whether its two chance-level cells contaminate that result. They do not: dropping the third seed moves the face perception rotation term by at most 0.0117 in absolute value at four of the five subject counts, and by +0.0675 at 20 subjects, meaning that including the seed makes the term **smaller** there. The one substantial shift runs against the factorial's conclusion rather than for it. A single collapsed fit inside a three-seed mean is invisible in the mean and visible in the spread, which is why S3.4 tabulates the per-seed spread beside every interval.

E.5 S5.5 Objection: the headline is a normalization-scope artifact

Each arm carries its own literature's normalization convention: per-recording robust scaling for the raw arm, per-frequency z-scoring fit on the pooled training draw for the spectral arms (S2.4, Section 3). The raw-versus-spectral contrast therefore bundles the representation with the scope at which its statistics are estimated.

The direction of the bias is not neutral, which is why this needed a control and not a sentence. A per-subject amplitude gain is an additive offset in log power. Per-recording robust scaling divides it out; per-frequency z-scoring fit on the pooled draw removes only the population mean of that offset and leaves the per-subject residual. On that account the raw arm is helped by its convention, and some unknown part of the headline gap is normalization rather than representation.

The control was pre-registered in full before any of its code existed. Two further arms match the scope in both directions rather than one:

- `R_pooled`: raw samples centered and scaled using per-channel median and interquartile range fitted on the pooled training draw, with clipping at 20 scaled units. This changes scope, not robust scaling into z-scoring.
- `M_recwave`: the spectral arm given the raw arm's convention, per-recording robust scaling of the waveform before the transform, followed by M's existing pooled per-frequency spectral z-score.

Running both matters: a single matched arm would leave the answer confounded with which arm was moved, and moving the arm expected to lose is the version of this control that is easiest to write and hardest to trust.

The reading was fixed before the runs. The confound is ruled out if, at n = 20, 40 and 80, the gap on both matched series keeps its direction and at least half its size, with the interval on that comparison excluding zero. Anything less is `PARTIAL`; a matched series that loses direction or more than half its size is `CONFOUND REAL`.

Result: the registered retention reading is met at every required level. Both matched-scope ERP gaps retain their direction and at least half the original magnitude, and every ERP interval on the primary unmatched and two matched-scope series excludes zero negatively. The headline runs 0.09 to 0.11 unmatched and 0.09 to 0.12 matched. Where one number is needed we quote the smaller of the two matched magnitudes, a convention fixed before the block ran so that the more favorable series could not be selected afterwards. This is not an equivalence test for an unchanged gap.

The mechanism we offered for this result is falsified, and we report that here rather than omitting it. The account above predicts that giving the raw arm the pooled convention costs it accuracy (`R_pooled` minus `R` negative) and that giving the spectral arm a per-recording convention gains it accuracy (`M_perrec` minus `M` positive). The pre-registered reading states that if those two effects come back with the same sign, the subject-nuisance account is wrong. They share a sign at n = 20, 160 and 240. The account is wrong, and whatever these arms measure is not what we said they measure. The verdict does not rest on the account: it was defined on the matched-scope comparison, which is unambiguous. But this control bought a clean answer to the objection while destroying the story we told about why the objection would fail.

A third arm, `M_perrec`, was pre-registered as secondary and stays secondary. It applies per-frequency per-recording z-scoring to the spectral arm, the more aggressive match, and moves the gap no further: -0.0924 to -0.1146 across the range. It does not enter the verdict, which was fixed on the two arms above before any of them ran.

What this does and does not settle. The measured controls rule out normalization scope as a sufficient explanation for the ERP deficit. They do not make the arms identical in normalization, which is not the goal. On the sleep corpus the same control leaves the oscillatory advantage in place at small n and above zero at n = 240 on both matched series, which is why the paper writes that the advantage decays toward zero rather than through it; the unmatched series alone would have licensed the stronger phrasing.

Stack. Every cell of this block reports one driver, one torch build and one CUDA runtime, so none of these comparisons crosses a stack boundary. That was a pre-registered precondition: the two scope effects are unquotable if it fails.

E.6 S5.6 Sleep token and compute control (C13-T)

The control tests whether the small-n sleep advantage depends on M's greater token count and computation. Direction A reduces M to R's 60 tokens; direction B increases R to M's 150 tokens (S2.1). All contrasts use the four arms rerun in this block; the unmatched n = 20 gap reproduces the archived advantage's direction before either matched contrast is interpreted.

The registered point-estimate reading is SURVIVES if both directions retain the within-block unmatched gap's direction and at least half its magnitude at n = 20 and 40, ARTIFACT if all four comparisons fail that criterion, and PARTIAL otherwise. Intervals are reported but do not determine the letter. The following are three-seed means with conditional 95 percent evaluation-person intervals, using zero as each contrast's null (E14).

n	M minus R	M-10f minus R (A)	M minus R-25f (B)
20	+0.0220 [+0.0093, +0.0342]	+0.0242 [+0.0111, +0.0358]	+0.0216 [+0.0069, +0.0366]
40	+0.0294 [+0.0128, +0.0450]	+0.0280 [+0.0157, +0.0403]	+0.0197 [+0.0031, +0.0351]
80	+0.0136 [+0.0019, +0.0246]	+0.0036 [-0.0087, +0.0149]	+0.0207 [+0.0105, +0.0301]
160	-0.0067 [-0.0173, +0.0024]	-0.0035 [-0.0140, +0.0055]	+0.0008 [-0.0098, +0.0116]
240	-0.0058 [-0.0176, +0.0055]	-0.0056 [-0.0156, +0.0046]	+0.0009 [-0.0118, +0.0132]

Both matched directions read SURVIVES at 20 and 40 subjects. Retention is 1.100 and 0.981 at n = 20, and 0.952 and 0.669 at n = 40, for A and B respectively. E14 assigns no verdict at the other three

counts; their contrasts and intervals are descriptive. At n = 80, direction A's interval includes zero; direction B stays weakly positive at n = 160 and 240. The advantage diminishes, but a crossover is not robust to both matches.

Per-seed maximum-minus-minimum spreads are shown separately; they are not confidence intervals and are absent from the conditional bootstrap uncertainty.

n	M minus R	M-10f minus R (A)	M minus R-25f (B)
20	0.0364	0.0395	0.0114
40	0.0097	0.0038	0.0160
80	0.0095	0.0141	0.0313
160	0.0048	0.0141	0.0076
240	0.0145	0.0096	0.0117

At n = 20, the unmatched and direction-A spreads, 0.0364 and 0.0395, exceed their mean gaps, +0.0220 and +0.0242. The matched intervals exclude zero at both registered counts, but that observation was not the verdict criterion. The result concerns one corpus, one model scale, three seeds and one-second windows. Matching also changes averaging resolution or raw patch length; SURVIVES establishes no prescription for conventional sleep staging, other corpora, montages, models or front-end pairs.

E.7 S5.7 Corrected normalization-control tables

G is M minus R; G pooled is M minus R-pooled; G recording is M-recwave minus R; G spectral is M-perrec minus R. The additional scope-effect series compare each changed arm with its original counterpart, as recorded in the report. The paired ERP changes do not support deepening. Both matched sleep changes are negative, while their endpoint intervals include zero and their point estimates remain positive.

Intervals condition on the fitted models, selected training draws and observed cohort membership counts. Seed spread is the reported maximum-minus-minimum of three fitted-seed points, not an interval.

Task	Contrast	Claim	n	Point	Conditional 95% CI	Seed spread
Evoked	G	C19	20	-0.0925	[-0.1098, -0.0750]	0.0198
Evoked	G	C19	40	-0.0909	[-0.1075, -0.0737]	0.0075
Evoked	G	C19	80	-0.1126	[-0.1316, -0.0948]	0.0229
Evoked	G	C19	160	-0.1096	[-0.1258, -0.0946]	0.0058
Evoked	G	C19	240	-0.1122	[-0.1273, -0.0966]	0.0131
Evoked	G pooled	C19	20	-0.1063	[-0.1234, -0.0892]	0.0194
Evoked	G pooled	C19	40	-0.0919	[-0.1084, -0.0745]	0.0284
Evoked	G pooled	C19	80	-0.1082	[-0.1248, -0.0906]	0.0153
Evoked	G pooled	C19	160	-0.1032	[-0.1192, -0.0878]	0.0143
Evoked	G pooled	C19	240	-0.1011	[-0.1169, -0.0844]	0.0217
Evoked	G recording	C19	20	-0.1027	[-0.1206, -0.0844]	0.0108
Evoked	G recording	C19	40	-0.0929	[-0.1096, -0.0755]	0.0188
Evoked	G recording	C19	80	-0.1048	[-0.1227, -0.0872]	0.0018
Evoked	G recording	C19	160	-0.1079	[-0.1243, -0.0934]	0.0067
Evoked	G recording	C19	240	-0.1154	[-0.1309, -0.1011]	0.0156
Evoked	S raw	C19	20	+0.0138	[+0.0075, +0.0196]	0.0141
Evoked	S raw	C19	40	+0.0011	[-0.0055, +0.0077]	0.0209
Evoked	S raw	C19	80	-0.0044	[-0.0105, +0.0010]	0.0119
Evoked	S raw	C19	160	-0.0064	[-0.0139, +0.0003]	0.0085
Evoked	S raw	C19	240	-0.0111	[-0.0180, -0.0040]	0.0086
Evoked	S wave	C19	20	-0.0102	[-0.0166, -0.0035]	0.0150
Evoked	S wave	C19	40	-0.0021	[-0.0074, +0.0033]	0.0166
Evoked	S wave	C19	80	+0.0078	[+0.0021, +0.0140]	0.0222
Evoked	S wave	C19	160	+0.0017	[-0.0042, +0.0080]	0.0022

Task	Contrast	Claim	n	Point	Conditional 95% CI	Seed spread
Evoked	S wave	C19	240	-0.0032	[-0.0090, +0.0022]	0.0280
Evoked	G spectral	C19	20	-0.0924	[-0.1088, -0.0755]	0.0221
Evoked	G spectral	C19	40	-0.1038	[-0.1211, -0.0860]	0.0097
Evoked	G spectral	C19	80	-0.1104	[-0.1282, -0.0935]	0.0062
Evoked	G spectral	C19	160	-0.1120	[-0.1278, -0.0973]	0.0088
Evoked	G spectral	C19	240	-0.1146	[-0.1303, -0.0989]	0.0013
Evoked	S spectral	C19	20	+0.0000	[-0.0059, +0.0060]	0.0160
Evoked	S spectral	C19	40	-0.0129	[-0.0197, -0.0061]	0.0094
Evoked	S spectral	C19	80	+0.0022	[-0.0037, +0.0083]	0.0208
Evoked	S spectral	C19	160	-0.0024	[-0.0091, +0.0042]	0.0146
Evoked	S spectral	C19	240	-0.0024	[-0.0083, +0.0037]	0.0131
Sleep	G	C19	20	+0.0220	[+0.0093, +0.0342]	0.0364
Sleep	G	C19	40	+0.0294	[+0.0128, +0.0450]	0.0097
Sleep	G	C19	80	+0.0136	[+0.0019, +0.0246]	0.0095
Sleep	G	C19	160	-0.0068	[-0.0173, +0.0023]	0.0048
Sleep	G	C19	240	-0.0060	[-0.0175, +0.0050]	0.0164
Sleep	G pooled	C19	20	+0.0436	[+0.0303, +0.0569]	0.0292
Sleep	G pooled	C19	40	+0.0363	[+0.0183, +0.0517]	0.0141
Sleep	G pooled	C19	80	+0.0401	[+0.0285, +0.0508]	0.0131
Sleep	G pooled	C19	160	+0.0164	[+0.0076, +0.0250]	0.0204
Sleep	G pooled	C19	240	+0.0056	[-0.0042, +0.0155]	0.0174
Sleep	G recording	C19	20	+0.0401	[+0.0258, +0.0538]	0.0297
Sleep	G recording	C19	40	+0.0419	[+0.0250, +0.0575]	0.0092
Sleep	G recording	C19	80	+0.0194	[+0.0076, +0.0308]	0.0059
Sleep	G recording	C19	160	+0.0072	[-0.0046, +0.0181]	0.0074
Sleep	G recording	C19	240	+0.0021	[-0.0074, +0.0121]	0.0115
Sleep	S raw	C19	20	-0.0216	[-0.0309, -0.0120]	0.0080
Sleep	S raw	C19	40	-0.0069	[-0.0162, +0.0026]	0.0091
Sleep	S raw	C19	80	-0.0265	[-0.0369, -0.0173]	0.0071
Sleep	S raw	C19	160	-0.0232	[-0.0314, -0.0153]	0.0155
Sleep	S raw	C19	240	-0.0116	[-0.0183, -0.0050]	0.0080
Sleep	S wave	C19	20	+0.0181	[+0.0070, +0.0291]	0.0111
Sleep	S wave	C19	40	+0.0125	[+0.0040, +0.0203]	0.0018
Sleep	S wave	C19	80	+0.0057	[-0.0032, +0.0135]	0.0120
Sleep	S wave	C19	160	+0.0139	[+0.0046, +0.0223]	0.0122
Sleep	S wave	C19	240	+0.0080	[+0.0001, +0.0169]	0.0048
Sleep	G spectral	C19	20	+0.0500	[+0.0356, +0.0650]	0.0216
Sleep	G spectral	C19	40	+0.0600	[+0.0461, +0.0750]	0.0134
Sleep	G spectral	C19	80	+0.0432	[+0.0312, +0.0550]	0.0079
Sleep	G spectral	C19	160	+0.0300	[+0.0211, +0.0399]	0.0041
Sleep	G spectral	C19	240	+0.0225	[+0.0124, +0.0323]	0.0111
Sleep	S spectral	C19	20	+0.0281	[+0.0156, +0.0413]	0.0178
Sleep	S spectral	C19	40	+0.0306	[+0.0182, +0.0442]	0.0081
Sleep	S spectral	C19	80	+0.0296	[+0.0193, +0.0407]	0.0080
Sleep	S spectral	C19	160	+0.0367	[+0.0263, +0.0480]	0.0059
Sleep	S spectral	C19	240	+0.0285	[+0.0156, +0.0426]	0.0053

Task	Contrast	Claim	Paired contrast	Point	Conditional 95% CI	Seed spread
ERP	n = 240 minus n = 20	C19	G_pooled	+0.0052	[-0.0054, +0.0159]	not reported
ERP	n = 240 minus n = 20	C19	G_recording	-0.0127	[-0.0266, +0.0006]	not reported
Sleep	n = 240 minus n = 20	C19	G_pooled	-0.0380	[-0.0530, -0.0218]	not reported
Sleep	n = 240 minus n = 20	C19	G_recording	-0.0380	[-0.0494, -0.0252]	not reported

F S6. Architecture controls and model input categories

F.1 S6.1 Intervention and scope

A learned linear projection can absorb a fixed frequency permutation into its weights. The original flat-embedding permutation control therefore does not identify the value of frequency order. The convolutional block tests log-power, complex and permuted-log-power inputs with an embedding that processes neighboring time-frequency bins. It contains 135 cells: three arms, three paradigms, five subject counts and three fitted seeds.

The embedding applies a (5, 3) convolution over frequency and time, GELU, pooling into four frequency bands, projection to model width, and normalization. Complex inputs supply real and imaginary coefficients as two input channels. Token sequences, positional embeddings, transformer and classifier stay fixed. This tests one embedding family, not an entire published pretrained model. Brant-2 processes the time-frequency plane convolutionally (Yuan et al., 2024); BrainBERT instead uses a linear spectral-column map (Wang et al., 2023).

The log-power embedding has 3,200 parameters in M and 4,352 in M-conv. The complex embedding has 6,272 in M-phi and 4,472 in M-phi-conv. Whole-model parameter counts remain close, but neither these counts nor the matched training schedule imply equal computation (Section 3).

F.2 S6.2 Frequency order matters without rescuing log power (C15, C17)

Permuting frequency reduces the convolutional arm's accuracy at every tested subject count. The ERP ordering contrast ranges from +0.0065 to +0.0153, with corrected conditional intervals excluding zero throughout. The committed resolution audit retains the registered C17 reading at all three resampling budgets (S3). This establishes sensitivity to frequency order within the tested convolution, not a general advantage of convolution over a linear projection.

M-conv remains below raw throughout, from -0.0881 at n = 20 to -0.1083 at n = 240. Its difference from M has no positive interval exclusion. Thus the registered C15 reading extends the raw-versus-log-power result to this frequency-structured embedding. The historical status label `L2 retired` applies to that tested objection, not to every possible frequency-aware architecture.

The sleep convolutional point series changes sign over the same adjacent levels as the original linear series. This descriptive agreement does not establish a normalization-robust crossover: the matched-scope sleep curves remain positive at the final level and their endpoint intervals include zero (S5.5).

F.3 S6.3 The original design prescription fails (C16, C18)

C16 historically described a `lossless`, frequency-structured design and predicted that it would beat both neighboring designs. Neither the claimed losslessness nor the supposedly empty literature category was established. The operational test is nevertheless clear: M-phi-conv must exceed both M-conv and M-phi, with positive intervals at every registered level.

The complex-input convolution beats its log-power counterpart, but does not beat the linear complex-input arm under that reading. Its embedding contrast has negative exclusions at n = 40 and n = 240, and no positive exclusion. C16 remains not supported, and its universal design prescription is withdrawn.

The narrow complex convolution has fewer embedding parameters than its linear comparator. C18 tests that objection with six pooled bands and 6,520 embedding parameters, exceeding the linear comparator's 6,272. The wider arm still has no positive interval exclusion against the linear arm; its n = 80 contrast excludes zero negatively. The registered capacity-control reading is met. This shows that increasing capacity in this particular manner does not rescue the prediction, rather than identifying architecture as the unique cause or establishing equivalence among embeddings.

F.4 S6.4 Collapsed fits and training variability

The original complex linear arm has one collapsed N170 fit at n = 20; the narrow convolution adds two in the same paradigm at n = 20 and n = 40. The wider convolution adds another. All fits remain in the primary estimates. A collapsed fit can strongly affect an average over three seeds, especially the small-sample complex-input contrasts. The seed spreads in the tables expose that sensitivity without claiming that conditional subject intervals cover it.

Two chance-level rotated-complex cells are not automatically collapsed fits: their broader arm series also performs poorly. The deliberate rotation changes cross-frequency relationships and alignment, so a low score has a different interpretation from a lone failed fit whose sibling seeds learn successfully. The historical failed implementation and its correction are retained in S4.4.

F.5 S6.5 Corrected architecture tables

The table labels use operational input and embedding contrasts. The immutable report key `C16_losslessness` is retained only for provenance; its presence in an old schema does not establish invertibility.

Intervals condition on the fitted models, selected training draws and observed cohort membership counts. Seed spread is the reported maximum-minus-minimum of three fitted-seed points, not an interval.

Task	Contrast	Claim	n	Point	Conditional 95% CI	Seed spread
ERP	M-conv minus R	C15	20	-0.0881	[-0.1073, -0.0696]	0.0115
ERP	M-conv minus R	C15	40	-0.0964	[-0.1138, -0.0781]	0.0077
ERP	M-conv minus R	C15	80	-0.1099	[-0.1290, -0.0910]	0.0043
ERP	M-conv minus R	C15	160	-0.1118	[-0.1275, -0.0975]	0.0117
ERP	M-conv minus R	C15	240	-0.1083	[-0.1241, -0.0925]	0.0120
ERP	M-conv minus M	–	20	+0.0044	[-0.0030, +0.0111]	0.0083
ERP	M-conv minus M	–	40	-0.0056	[-0.0125, +0.0013]	0.0130
ERP	M-conv minus M	–	80	+0.0027	[-0.0025, +0.0080]	0.0272
ERP	M-conv minus M	–	160	-0.0022	[-0.0091, +0.0047]	0.0172
ERP	M-conv minus M	–	240	+0.0039	[-0.0019, +0.0092]	0.0168
ERP	M-phi-conv minus M-conv	C16	20	+0.0482	[+0.0364, +0.0597]	0.0936
ERP	M-phi-conv minus M-conv	C16	40	+0.0426	[+0.0313, +0.0539]	0.0987
ERP	M-phi-conv minus M-conv	C16	80	+0.0938	[+0.0809, +0.1071]	0.0127
ERP	M-phi-conv minus M-conv	C16	160	+0.0932	[+0.0795, +0.1067]	0.0098
ERP	M-phi-conv minus M-conv	C16	240	+0.0804	[+0.0686, +0.0916]	0.0298
ERP	M-phi-conv minus M-phi	C16	20	-0.0027	[-0.0092, +0.0039]	0.0191
ERP	M-phi-conv minus M-phi	C16	40	-0.0424	[-0.0498, -0.0347]	0.0886
ERP	M-phi-conv minus M-phi	C16	80	-0.0018	[-0.0091, +0.0063]	0.0382
ERP	M-phi-conv minus M-phi	C16	160	+0.0015	[-0.0042, +0.0071]	0.0042
ERP	M-phi-conv minus M-phi	C16	240	-0.0099	[-0.0163, -0.0034]	0.0073
ERP	M-conv minus permuted	C17	20	+0.0131	[+0.0072, +0.0187]	0.0156
ERP	M-conv minus permuted	C17	40	+0.0113	[+0.0038, +0.0189]	0.0094
ERP	M-conv minus permuted	C17	80	+0.0065	[+0.0009, +0.0123]	0.0062
ERP	M-conv minus permuted	C17	160	+0.0125	[+0.0073, +0.0177]	0.0061
ERP	M-conv minus permuted	C17	240	+0.0153	[+0.0092, +0.0213]	0.0133
Sleep	M-conv minus R	C15	20	+0.0155	[+0.0042, +0.0288]	0.0323
Sleep	M-conv minus R	C15	40	+0.0245	[+0.0123, +0.0368]	0.0191
Sleep	M-conv minus R	C15	80	+0.0087	[-0.0040, +0.0201]	0.0118
Sleep	M-conv minus R	C15	160	-0.0009	[-0.0134, +0.0093]	0.0132
Sleep	M-conv minus R	C15	240	-0.0119	[-0.0211, -0.0022]	0.0108
ERP	wide conv minus M-phi	C18	20	-0.0010	[-0.0081, +0.0059]	0.0397
ERP	wide conv minus M-phi	C18	40	-0.0048	[-0.0110, +0.0014]	0.0557
ERP	wide conv minus M-phi	C18	80	-0.0100	[-0.0168, -0.0032]	0.0273
ERP	wide conv minus M-phi	C18	160	+0.0006	[-0.0055, +0.0066]	0.0099
ERP	wide conv minus M-phi	C18	240	-0.0030	[-0.0093, +0.0033]	0.0135
ERP	wide minus narrow conv	C18	20	+0.0017	[-0.0038, +0.0071]	0.0205
ERP	wide minus narrow conv	C18	40	+0.0376	[+0.0287, +0.0472]	0.1209
ERP	wide minus narrow conv	C18	80	-0.0081	[-0.0144, -0.0018]	0.0109

Model	Observed input	Comparison
CBraMod	Raw + FFT energy	Branch removal
TFM-Tokenizer	Raw / STFT / both	Input representation
BrainBERT	STFT / superlet	Spectral transform
LaBraM	Raw patches	Prediction target
BandVQ	Band waveforms	Auxiliary tokens
This study	Raw / power / complex	Input operations and embedding

Figure 5: **Input, transform, target, and auxiliary-token ablations answer different questions.** Observable inputs and comparisons replace the unproved lossless/empty-category taxonomy. The original prescription is withdrawn; its tested C16 conjunction remains negative.

Task	Contrast	Claim	n	Point	Conditional 95% CI	Seed spread
ERP	wide minus narrow conv	C18	160	-0.0009	[-0.0079, +0.0056]	0.0057
ERP	wide minus narrow conv	C18	240	+0.0068	[+0.0017, +0.0118]	0.0062

F.6 S6.6 Observable model properties and the complete ERP arm ladder

Figure 5 records input operations and comparison categories, without inferring invertibility or total phase loss from summation, vector quantization, or convolution. CBraMod directly ablates temporal and frequency branches; TFM-Tokenizer compares raw, STFT and combined inputs (Wang et al., 2024b; Pradeepkumar et al., 2026). The figure's other rows are characterized from their own reports: raw-patch encoders (Ouahidi et al., 2025; Wang et al., 2024a), an intracranial time-frequency model (Zhang et al., 2023; Yuan et al., 2024), and band-wise auxiliary tokens (Sukhbaatar et al., 2026). These focused corrections do not constitute a complete re-audit of the recorded model catalog, so the catalog supports no prevalence or absence count. Figure 6 retains every original input arm alongside the focused paired-gain comparison in main-text Figure 4.

G S7. Evidence index and reproducibility boundaries

G.1 S7.1 Anonymous evidence identifiers

The identifiers below map manuscript comparisons to preserved run metadata. They do not imply that the archive has been publicly released. Archive A is the immutable post-review snapshot, identified by the SHA-256 of its manifest:

`91ccd65cca9e63b50b2734b7ac4fce2b2a0f7a02aa35bb383ea452367f2d6095`.

Bundle names below are relative to A's run directory. Within each bundle, the complete stable run ID selects its metrics and, where retained, evaluation predictions. A run ID combines dataset, paradigm, arm, model size, subject count, seed and configuration hash; it contains no individual participant identifier. The private machine-readable index `evidence-index.json` maps every listed run to its full fitted-code revision, configuration hash, data-manifest hash and metrics-file checksum. Its rendered summary follows. A data-manifest hash identifies the recorded input selection, not an upstream dataset release or a recoverable list of raw-file checksums.

E01: Primary ERP/sleep sweeps; C1–C3, C11–C13.

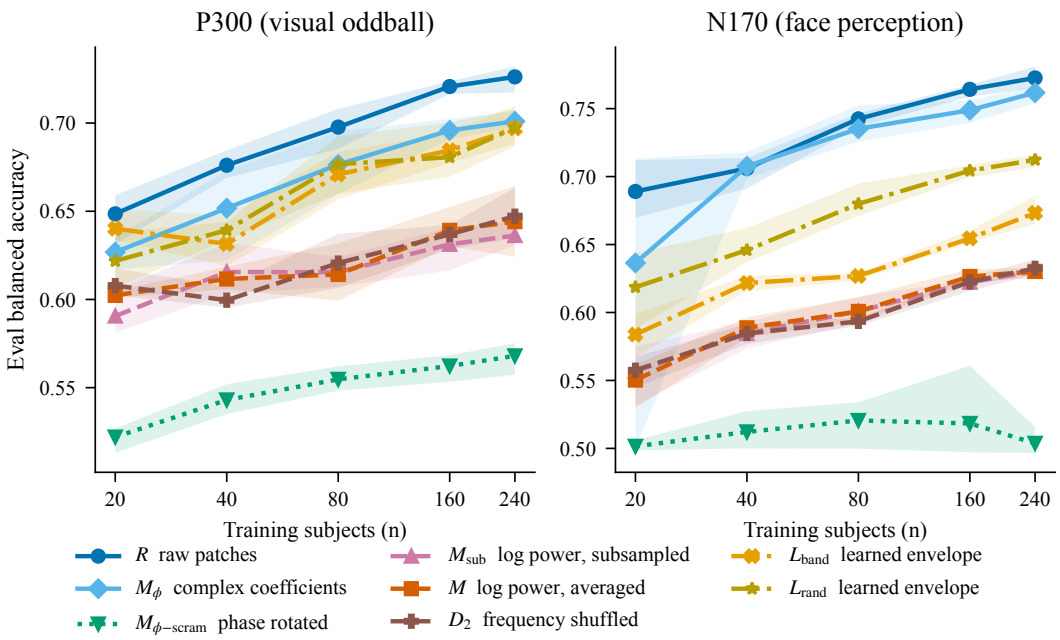

Figure 6: **Learned envelopes qualify the phase-based grouping, especially on P300.** Lines show mean evaluation balanced accuracy over three fitted seeds; shading is the seed minimum-to-maximum range, not a confidence interval. On N170 the complex and raw arms lead the envelope arms throughout; on P300 learned envelopes approach the complex arm and exceed it at n = 20 and n = 80. M-phi improves over M without establishing equivalence to raw or universal separation of all complex and envelope arms.

Archive bundles: `sweep`. Indexed runs: 285. Recorded fitted-code revisions: 23f373c3, 297a9692, 34b7f81c, 450ff393, 46139321, 8276103d, 92638477, a0205100, af3b1832, b820f506, bfd587be. Example stable run ID: `nm000225_sleep_M_M_n160_seed1_2a1c4e804939`.

E02: Convolutional controls; C15–C17.

Archive bundles: `sweep_conv`. Indexed runs: 135. Recorded fitted-code revisions: 923a534d. Example stable run ID: `nm000225_sleep_D2_conv_M_n160_seed1_a1c5e3e5c463`.

E03: Wide convolution; C18.

Archive bundles: `sweep_conv_wide`. Indexed runs: 30. Recorded fitted-code revisions: 182a90ce-dirty, 26f61d99, 8bec44d8-dirty, 9e9eedcf. Example stable run ID: `on007056_p300_M_phi_conv_wide_M_n160_seed1_d4d959d10c66`.

E04: Normalization-scope controls; C19.

Archive bundles: `sweep_scope`. Indexed runs: 135. Recorded fitted-code revisions: 383d43e4, 64def518, b982b17c, e2532b42. Example stable run ID: `nm000225_sleep_M_perrec_M_n160_seed1_e5b2944bea36`.

E05: Historical SSL transfer and budget control; C4.

Archive bundles: `ssl`; `ssl-transfer`; `ssl-transfer-budget-matched`. Indexed runs: 63. Recorded fitted-code revisions: 909c8e51, af3b1832, d4d8424b. Example stable run ID: `ssl_M_M_steps100000_batch128_seed1_6c6a997063ca`.

E06: Archived saved-prediction SSL reruns; C4, superseded for the interval by E12.

Archive bundles: `ssl-transfer-with-predictions`; `ssl-transfer-budget-matched-with-predictions`. Indexed runs: 60. Recorded

fitted-code revisions: b982b17c. Example stable run ID: `on007056_p300_ssl_M_M_fine_tune_n20_seed1_80d0066ad859`.

E08: Six-channel P300 control.

Archive bundles: `erp-6ch`. Indexed runs: 30. Recorded fitted-code revisions: 17791c01. Example stable run ID: `on007056-6ch_p300_M_M_n160_seed1_6deacb2ecb66`.

E10: Validation and collapsed-grid diagnostics.

Archive bundles: `validate`; `collapsed-grid-diagnostic`. Indexed runs: 6. Recorded fitted-code revisions: 7c84eaff, 80bf4f90-dirty, a0205100, de8f6e36-dirty. Example stable run ID: `on007056_p300_M_phi_conv_M_n20_seed1_29cecc47787f`.

G.2 S7.2 Analysis versions and result mapping

Analysis bundle B is the reviewed subject-cluster revision at code `c07020e51a04b27f4262e0a1841d998faca445f0`, following amendment 8780c13. Its audit links every corrected report to A's manifest, the original report checksum, corrected report checksum and analysis-file checksums. This is an analysis revision, not the fitted-model version. The primary reports are:

Evidence / manuscript result	Report identifier in B
E01; ERP gap, C1	`erp-M-R.json`
E01; complex versus log power, C11	`erp-Mphi-M.json`
E01; rotation, C12	`erp-Mphi-Mphiscram.json`
E01; residual representation, C12	`erp-Mphiscram-Msub.json`
E01; averaging, C12	`erp-Msub-M.json`
E01; sleep contrasts, C13	`sleep-M-R.json`; `sleep-Mphi-M.json`
E02; architecture and ordering, C15–C17	`conv-verdicts-reread.json`; `c17-resolution.json`
E03; capacity, C18	`c18-verdict.json`
E04; normalization scope, C19	`c19-verdict.json`
E06; archived C4 reruns, diagnostic only, superseded by E12	`c4-bootstrap.json`

The full arm ladder, learned-filter comparisons and six-channel control use the evaluation metrics in E01 and E08 directly. Their descriptive readout is in the manuscript claim checker at revision 8ec7d50; this does not substitute for the fitted-code revisions above.

E07: published-checkpoint probes (C5, C6). The stable report IDs are `probe-regression-reve-c89d47f791e8` and `probe-regression-labram-6cc828fc535c`. Both record probe code revision `cddebdf0e1a4a3a999a99b2260b001c23c5c40b0`, input-population checksums, checkpoint checksums and ridge settings. S2.9 gives the checkpoint/input contracts. The published models' upstream pretraining-code versions are not reconstructed from our probe commit.

E09: historical cross-stack comparison (C10). `cross-stack.json` preserves the 54 paired cell descriptions and `a1-cross-stack.md` the original formatted comparison. The comparison artifact does not record source-run IDs or fitted/analysis commits sufficient to reconstruct the other stack from A. This is a provenance gap; the table's saved point differences are not an independently reproducible cross-stack run collection.

E10: implementation validation and scale diagnostic. The stable corrected validation run is `on007056_p300_M_phi_conv_M_n20_seed1_29cecc47787f`. The reconstructed scale diagnostic is `diagnose_conv_scale.py`, as preserved at manuscript revision 8ec7d50. The original failed run was overwritten; its output cannot be recovered.

E11: historical linear-accessibility exploration (C14). A contains `reconstruction/recon-on007056.json`, `reconstruction/recon-on007096.json`,

reconstruction/recon-nm000225.json and the corresponding nuisance-v2 reports. These summaries do not record enough code and input-version metadata for an exact replay. They preserve the historical failed predictor, not confirmatory evidence for a front-end rule.

E12: deterministic C4 re-run and six-seed interval (C4). Archive C is a separate snapshot of one block, identified by the SHA-256 of its manifest:

06e21ba63ea4c677a227500cabfef2a00b8503cd17fe3d4d7f6644e72878e31a.

Its runs share fitted-code revision efe53d4a6479a94cf81b58cf235bf5342b32c8a5 and one node allocation, and each records its device, its deterministic-algorithm setting and its software stack. Example stable run ID: on007056_p300_ssl_M_M_fine_tune_n20_seed1_d88f8bb83668. Its two read-outs, each pinned to that manifest, are identified by method: c4-e2-deterministic-rerun-v1 for the exact-repeat and device readings, and c4-e3-seed-interval-v1 for the six-seed interval, a seed interval that carries a device component.

E13: native-objective premise check (C4-E4P). The read-out c4-e4p-premise-v2, at analysis code revision 92c33f10c91aed6a4396ec47ae6407de703716c2, evaluates the three native pretrained models of E05, ssl_R_M_steps100000_batch128_seed1_7052d4c4bedb, ssl_M_M_steps100000_batch128_seed1_6c6a997063ca and ssl_M_phi_M_steps100000_batch128_seed1_0eb78e74a521, on epochs drawn from the pretraining pool identified by manifest hash 84355faea527bbcb824922e3831e3a83061d1c9c188efa735df394641ce9ea1b. It reads no evaluation population.

E14: sleep token and compute control (C13-T). Archive D contains the 60 deterministic sleep cells and their predictions, identified by manifest SHA-256:

c83620b2b6017e92b183800c8e14dc7e8d49a762de94653eb21a233454d48266.

All fitted cells use revision 8746f80239b73abff76958303f1205053396ab77. The aggregate read-out is 2026-09-16-phase4-sleep-token-control.json, method c13t-sleep-token-control-v1; it verifies 121 archived artifacts and carries the full curve, conditional intervals, seed spreads and the point-retention reading. The matched comparisons use only D's own rerun unmatched arms (S2.1, S5.6).

E15: rotation-label diagnostic (P2-A).

The aggregate record is:

2026-09-16-phase2-rotation-label.json.

Method phase2-rotation-label-independence-v1, records the diagnostic at reader revision 727da31. It uses 600 real P300 epochs from three local sample people, matching preprocessing hash 9809f8bc1194 and the 29-channel layout to manifest-verified archive A. These are not the production runs' angle tensors. Its inconclusive resultant-length test is specified in S2.3.

E16: common-target pretraining (C4-E4). Archive E contains the three completed replicates, identified by manifest SHA-256:

1aac6f9c47b489a45bb4d78af2e6c941492b2cb452a13f4ba7b640868c810fc2.

All 162 executions share clean deployed revision:

8c18fd32da1658806235ffe6ab92fcfe23166b47.

The aggregate is 2026-09-22-phase6-e4-common-target.json, method c4-e4-common-target-v1; its reader verifies the 327 metric, prediction, checkpoint-sidecar and block-provenance artifacts it consumes. The preserved snapshot has 330 files. The reader at revision f3711457b6c9430c9c1938c20cfd99704634fea8 was populated with the manifest pin above after preservation; its exact bytes and the aggregate checksum are recorded in 2026-09-25-manuscript-evidence-map.json. S2.8, S3.3 and S5.3 distinguish its design, seed interval with node/device components, manipulation checks and E4-NULL reading from native-target E12.

The manuscript checkout carries the aggregate records for E14–E16 and their checksum map. It does not carry their raw snapshots or all audit dependencies; those remain at the recorded experiment revisions and preserved locations. Reading these summaries is not replaying the raw analyses or retraining the models. The records and their code provenance remain private.

G.3 S7.3 Limits of reproducibility and availability

Some fitted records carry a dirty-tree marker, including wide-convolution and diagnostic runs. The indexed base commit does not identify an unrecorded working-tree delta. Where no matching patch was archived, exact reconstruction of that code state remains unavailable. The input manifest hashes likewise do not recover missing raw-source release identifiers (S2.5). These limitations are reported separately from the measured evaluation results.

The archived C4 reruns (E06) differ from nominally matching executions, and their stack metadata coverage is incomplete; the audit did not identify why the drift concentrated in M-phi. The reported C4 numbers come instead from archive C (E12), whose runs record their stack and device and whose same-device repeats agree exactly.

The code, evidence index and archive currently reside in a private repository. An anonymous public artifact has not been released or verified. This rendered index makes evidence identifiers visible without exposing individual participant IDs, account names or identifying host paths. It does not claim that a reviewer can currently download the private archive. Publication permission, de-identification and an accessible anonymous export remain separate submission-readiness requirements. Third-party EEG data are not redistributed.