# LEARNING EEG STRUCTURE IN ITS NATIVE FEATURE SPACES: OPERATOR-PRECONDITIONED MULTI-VIEW PRETRAINING

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

EEG foundation models are trained to learn general-purpose latent representations, but different physiological structures are not equally easy to discover in the same feature space of the signal. We study feature-space mismatch: under finite data, model capacity, optimization, and readout budgets, a structure may be present in EEG while remaining difficult for a learned representation to expose and use. A deterministic change of feature space does not add information, but can make such structure more accessible to a finite learner. Motivated by this observation, we propose operator preconditioning: explicitly compute structure-aware EEG measurements, then use self-supervision to learn their organization in the resulting feature space. A learning-free operator already reads a known physiological change more accurately than the frontends, and even the pretrained encoders, of the EEG foundation models we evaluate, highlighting the importance of the feature space before representation learning itself. We instantiate this principle as OpView, a multi-view EEG foundation model with independently pretrained waveform, spectral, cross-spectral, spatial, and dynamical views. Operators can be replaced or extended according to the structures and invariances relevant to different task families. Using operator-only, matched random-encoder, and pretrained-encoder controls, we separate the contributions of measurement, architecture, and learning. Under the EEG-FM-Compass evaluation protocol (Liu et al., 2026b), OpView achieves strong performance across multiple EEG paradigms. These results support operator-preconditioned representation learning as an extensible approach for directing foundation-model capacity toward structures that downstream EEG decoding can effectively use.

## 1 INTRODUCTION

EEG has long been decoded through several views of the same signal: motor imagery through band power and the spatial covariance of band-limited signals (Pfurtscheller & Lopes da Silva, 1999; Blankertz et al., 2008; Ang et al., 2012; Barachant et al., 2012), a steady-state visual evoked potential (SSVEP) through the energy at the stimulation frequency and its harmonics (Lin et al., 2006; Nakanishi et al., 2018). Each view makes some activity explicit and hides the rest; induced activity, for instance, is clear in band power but cancels in the trial average (Pfurtscheller & Lopes da Silva, 1999). Decoders therefore choose the view to suit the paradigm (Lotte et al., 2018).

EEG foundation models take the opposite approach. They pretrain one representation for every task, mostly with one self-supervised objective on patches of time-domain samples, and leave the network to discover every structure a task may need (Jiang et al., 2024; Wang et al., 2025; 2024; Liu et al., 2026b). The approach has not yet paid off: specialist models trained from scratch remain competitive with twelve published foundation models on many tasks (Liu et al., 2026b).

We trace part of this shortfall to that reliance. How readily a finite learner discovers a structure depends on the *feature space* it is given, the space into which a fixed map sends the input and in which the learner compares inputs (Bishop, 2006; Schölkopf & Smola, 2002). A deterministic change of feature space adds no information (Cover & Thomas, 2006), yet it changes what a learner with finite data, capacity, optimisation and readout can learn and use (Xu et al., 2020). A structure can therefore



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.

be present in EEG and still be less readable from a learned representation than from a suitable fixed operator; we call this gap *feature-space mismatch*.

Released EEG foundation models show this mismatch (Section 3). A fixed band-power operator recovers a known power change more accurately than CBraMod and EEGPT, and on SSVEP four of five encoders decode 16 to 22 points worse than their patch embeddings. Ablations trace the gap to two causes: the feature space and the objective each change what the same budget learns.

If the feature space and the objective decide what is learned, both should be chosen for the structure to be learned. In *operator preconditioning*, a fixed operator maps EEG into an observation space that holds a family of candidate features for one kind of structure, such as all band powers of all sensors, and an encoder pretrained to predict hidden or future values in that space learns which of these features matter and how they are organised across sensors, frequency and time. An operator, its prediction task and its encoder form a *view*. Spectral inputs have been used in EEG pretraining before (Yang et al., 2023; Döner et al., 2025); we choose the observation space, its invariances and the objective together, and train each view independently, so that views can be added or replaced one at a time. Because a space that helps one task can discard what another needs, our instantiation, OpView, has five views: Waveform, Frequency, Relation, Projective and Dynamics (Figure 2).

We contribute *(i)* a definition of feature-space mismatch, an analysis of its two causes and evidence of it in released EEG foundation models; *(ii)* operator-preconditioned multi-view pretraining, instantiated as OpView; and *(iii)* an evaluation that reads each view as the benchmark reads a published model, beside two controls: its operator output, which measures what the observation space already offers, and a random encoder of the same architecture, which measures what the architecture contributes. Every encoder learns its pretraining task, pretraining improves on the random encoder in 16 of 18 matched single-view comparisons, the views divide the work according to their observation spaces, and a single view beats the best of twelve published foundation models in 10 of 36 cells.

## 2 Related work

**EEG foundation models.** EEG foundation models follow language and vision (Devlin et al., 2019; He et al., 2022) in pretraining one representation with one self-supervised objective, usually on time-domain patches, by contrast, masked reconstruction, prediction of quantised codes or latent features, or forecasting (Kostas et al., 2021; Jiang et al., 2024; Wang et al., 2025; 2024; Cui et al., 2024). Benchmarks find that neither linear probing nor scale helps uniformly and that task-specific decoders remain strong (Liu et al., 2026b; Xiong et al., 2025; Schirrmeister et al., 2017; Lawhern et al., 2018; Lotte et al., 2018; Roy et al., 2019; Yang et al., 2026; Banville et al., 2021). The closest designs feed a transform of the signal to one encoder trained with one objective, as BIOT does with short-time spectra (Yang et al., 2023) and LUNA with Fourier magnitude and phase added to temporal patches (Döner et al., 2025) (Appendix I), while multi-view self-supervision learns what several views share (Tian et al., 2020; Federici et al., 2020; Tsai et al., 2021; Kumar et al., 2022b). Our views have their own operators, objectives and encoders, and are trained independently.

**Classical EEG features and preconditioning.** Our operators come from classical EEG analysis, from time-locked averages and band power to spatial covariance, coherency and polarity-free microstates (Luck, 2014; Pfurtscheller & Lopes da Silva, 1999; Barachant et al., 2012; Nolte et al., 2004; Michel & Koenig, 2018). A classical pipeline computes one chosen feature and stops; we keep the whole output as the observation space in which an encoder learns which features matter. Preconditioning is a term from optimisation, and input representations set how fast functions are learned (Nocedal & Wright, 2006; Jacot et al., 2018; Rahaman et al., 2019).

**Measurement tools.** To judge a learned view, we measure what it offers a readout, with usable information (Xu et al., 2020), probes (Alain & Bengio, 2017; Hewitt & Liang, 2019; Kumar et al., 2022a) and audits of EEG foundation models (Tang et al., 2026; Bindra & Panwar, 2026). Networks favour simple predictive features (Shah et al., 2020; Geirhos et al., 2020; Pezeshki et al., 2021), reconstruction rewards high-variance components (Balestriero & LeCun, 2024), and random features are strong baselines (Rahimi & Recht, 2007; Saxe et al., 2011); hence the two controls of every view.

# 3 FEATURE-SPACE MISMATCH LIMITS THE READABILITY OF FEATURES

EEG foundation models do not yet hold a clear advantage over decoders trained for a single task (Liu et al., 2026b). This section argues that feature-space mismatch is one cause: a model pretrained in an unsuitable feature space can leave structures that tasks rely on less readable than a fixed operator makes them. We define readability, state two factors of pretraining that affect it, test released models for mismatch, and vary each factor in small ablations.

## 3.1 READABILITY AND MISMATCH

We first define readability. Let $X$ be an EEG segment, $S = s(X)$ a structure such as a band power or a task label, and $Z = E(X)$ a representation. A readout protocol $b$ fixes the model family, the labelled samples, the preprocessing and the fitting rule; with $\widehat{h}_{b,\mathcal{D}}$ fitted on training data $\mathcal{D}$, readability is the expected readout error on independent test data,

$$\mathcal{R}_b(S \mid Z) = \mathbb{E}_{\mathcal{D}}\mathbb{E}_{X,S}\big[\ell\big(\widehat{h}_{b,\mathcal{D}}(E(X)), S\big)\big], \tag{1}$$

lower meaning more readable. The feature space alone can decide readability: when both polarities of the signal are equally likely, no affine function of the samples predicts the power $\|BX\|^2$ of a band-pass filter $B$ better than its mean, whereas an affine function of the power operator recovers it exactly (Appendix A.1). We speak of *feature-space mismatch* when $S$ is present in $X$ yet the same readout reads it worse from a learned representation than from a suitable fixed operator.

## 3.2 THE OBJECTIVE AND THE FEATURE SPACE

Mismatch is an outcome of pretraining, which makes a structure readable only if using the structure lowers the pretraining loss and the model can reach it from its feature space. To make the first condition precise, let $U$ be the visible context, $T$ the prediction target, $Q = q(U)$ the information the model already uses and $A = a(U)$ an additional structure, such as the band power of the visible patches. Under squared loss, using $A$ lowers the loss by at most its *reward*

$$\Delta_T(A \mid Q) = \inf_g \mathbb{E}\|T - g(Q)\|^2 - \inf_g \mathbb{E}\|T - g(Q, A)\|^2, \tag{2}$$

with infima over square-integrable measurable predictors.

**Proposition 1** (Reward of a squared-loss objective)**.** *For square-integrable $T$, let $m_Q = \mathbb{E}[T \mid Q]$, $m_{QA} = \mathbb{E}[T \mid Q, A]$ and $\delta = m_{QA} - m_Q$. Then $\Delta_T(A \mid Q) = \mathbb{E}\|\delta\|^2$.*

Proposition 1 has two consequences for EEG. First, an objective rewards a structure only through its contribution to the optimal prediction (Lee et al., 2021): if the phase of a masked rhythm cannot be predicted from the context, its power leaves the optimal waveform prediction unchanged, so a waveform objective offers no reward for power, while a target that predicts power does (Appendix A.2). Second, the feature space limits how much of the reward a model collects and how fast: a model collects only the share $\rho_{\mathcal{S}} \in [0, 1]$ of the reward that lies in the subspace $\mathcal{S}$ of changes it can make, and gradient descent with a fixed kernel can leave the small-eigenvalue part of a target unlearned within a finite budget (Appendix A.3). Sections 3.3 and 3.4 test them in turn.

## 3.3 MISMATCH IN RELEASED MODELS

Three experiments test the first consequence in released models pretrained on waveform patches.

**A known power change.** A quantity with a known true value compares a model's frontend and its learned representation with a fixed operator directly. After the $\mu/\beta$ power of sensorimotor electrodes in 64 motor-imagery recordings is scaled by known factors, a subject-held-out ridge recovers the injected log-power ratio with $R^2$ of 0.99 from a fixed power operator, against 0.63 and 0.61 from the patch embeddings of CBraMod and EEGPT (Figure 1a,b). Pretraining raises their encoders to 0.93 and 0.90, from 0.43 and 0.30 at random initialisation, still below the operator, with paired intervals that exclude zero (Appendix B.2). The feature space thus sets the starting point before any learning, and learning recovers most but not all of the gap; this is the mismatch of Section 3.1.

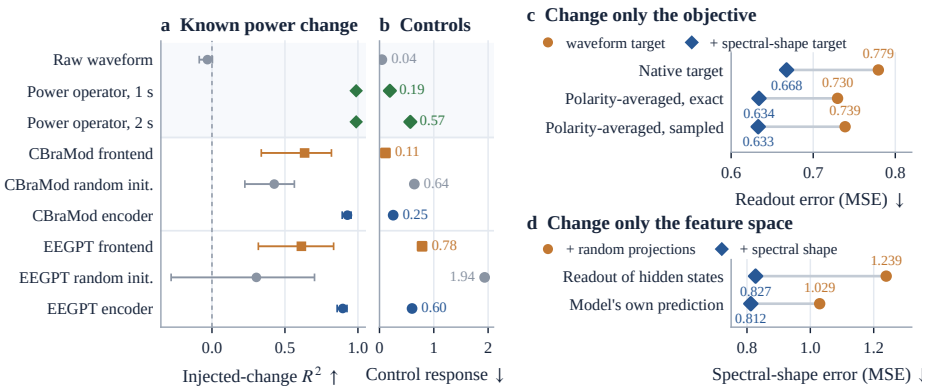



Figure 1: **The feature space and the objective each change what a model can read.** A fixed power operator recovers a known power change better than released encoders (a, b); adding a spectral-shape objective (c) or spectral-shape inputs (d) makes the spectral shape more readable at the same budget.

**Patch embeddings and encoder.** Learning can also make what the frontend exposes less readable. Across five released models on four datasets (Appendix B), the complete encoder decodes worse than its patch embeddings in 13 of 20 model–dataset pairs, and a randomly initialised encoder in 15, so much of this gap comes with the architecture. Pretraining narrows it in 14 pairs and widens it in 6, four of them on steady-state responses, where four of the five pretrained encoders separate the stimulation frequencies 16 to 22 points less well than their own patch embeddings: under this readout, pretraining makes the stimulation frequency less readable than the frontend leaves it, where classical decoders remain strong (Nakanishi et al., 2018; Liu et al., 2026b).

**Reconstruction and classification.** Classification can depend on what reconstruction ignores. Rotating the Fourier coefficients of all channels by one random phase per frequency preserves power and cross-spectra, yet lowers the SSVEP accuracy of frozen CBraMod and EEGPT under a fixed classifier from 55.6% to 13.3% and from 72.0% to 16.7%, while CBraMod's masked-reconstruction error does not rise (Appendix B.5).

### 3.4 Ablations of the feature space and the objective

Released models differ in both feature space and objective, so we changed each factor alone. To isolate the feature space, we trained one small masked-prediction model twice, differing only in four extra input values per patch, random projections of the waveform or its spectral shape; after 300 updates, its spectral-shape error, identical at the start, is 0.81 with the spectral shape and 1.03 with the projections (Figure 1d; Appendix B.4). To isolate the objective, we continued CBraMod for 64 steps with an added target, the spectral shape of masked patches; the error of reading that shape from its hidden states is 0.67, against 0.78 when the continuation keeps the waveform objective alone and 0.74 before it (Figure 1c; Appendices B.3 and B.6).

Feature-space mismatch thus limits the readability of features, and both the feature space and the objective contribute to it; Section 4 builds a method that chooses the two together.

## 4 Operator-preconditioned multi-view pretraining

Section 3 found that what a model learns depends on its feature space and on the objective posed there; operator preconditioning turns this into a design, stated in general and then for five views.

### 4.1 Framework

The method has three parts: a family of fixed operators $\{O_v\}_{v\in\mathcal{V}}$, one encoder per operator, and a task head (Figure 2),

$$Z_v = E_{\theta_v}\big(O_v(X)\big), \qquad \mathcal{L}_{\text{pre}} = \sum_{v\in\mathcal{V}} \mathcal{L}_v(\theta_v), \qquad \widehat{y} = h_\psi\big(\{Z_v\}_{v\in\mathcal{V}}\big). \tag{3}$$

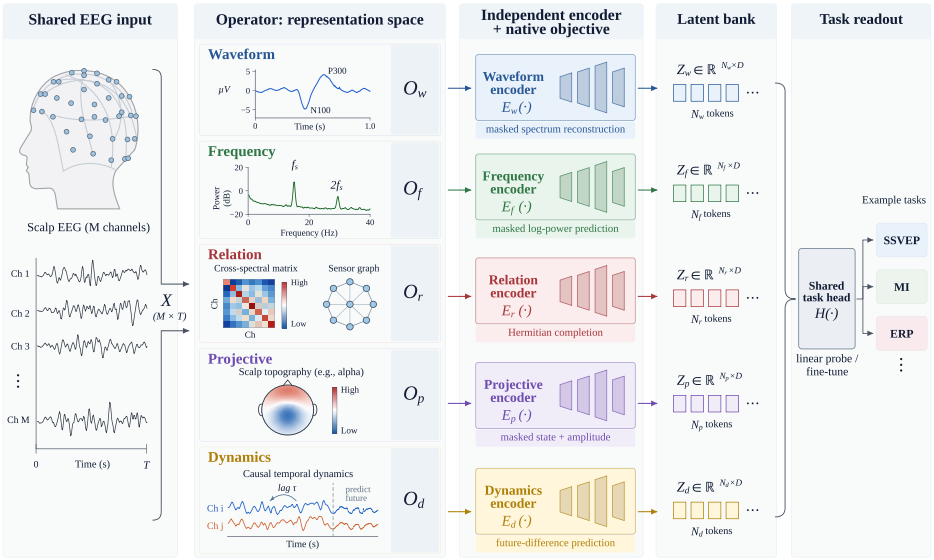



Figure 2: **Each view of OpView pairs a fixed operator with its own encoder and native objective.**

The operator sets the *observation space*: it fixes what is measured and which nuisance is removed, fitting at most normalisation statistics once on training data, and selects no single feature. Its output holds a family of candidate features; the log power of every sensor and 1 Hz group, for instance, contains every band power, band ratio and harmonic sum over any group of sensors as a simple function. The encoder learns which members of this family matter and how they are organised across sensors, frequency and time, and the head $h_\psi$ is fitted per task. Like a classical preconditioner, the operator is a fixed change of variables that makes a solution easier to reach, hence its name.

**Native objectives.** Section 3 suggests rewarding structure in the space where learning takes place, so the objective is defined on the operator output. With $\mathcal{M}_v$ the hidden or future positions of the operator output and $y_v$ a target computed there,

$$\mathcal{L}_v(\theta_v) = \mathbb{E}\,\ell_v\big(f_{\theta_v}\big(O_v(X)_{\setminus\mathcal{M}_v}\big),\, y_v\big) + \lambda R_v, \tag{4}$$

where $f_{\theta_v}$ is the encoder with a prediction head, $\ell_v$ a squared or Smooth-L1 error and $R_v$ a small auxiliary term. By Proposition 1, this objective rewards whatever in the visible context predicts the hidden observations, and since the masks act on the operator output, the target cannot be copied from the input. The objective is *native*: it is posed in the view's observation space, though input and target may differ in basis, as when the Waveform encoder reads samples and predicts their spectrum.

**Views and their invariances.** An operator, its native objective and its encoder form a *view*. Views share no parameters and are not aligned, so each can be replaced or added alone; only a combining head is refitted. No single view serves every task, because invariances cost information: if $O(gX) = O(X)$ for a nuisance action $g$, no head on $O(X)$ can separate labels that differ only along the orbit of $g$ (Appendix A.4). An invariance thus helps tasks that share it and can hurt the others (Cohen & Welling, 2016; Bronstein et al., 2021), and the waveform, which keeps everything, can still leave a structure hard to learn (Section 3); the method therefore learns in several observation spaces.

**Designing a view.** The same reasoning gives a recipe for adding a view: name the family of structures and the observations it needs; choose an operator whose output space holds that family, stating what it keeps and removes; pose a native prediction task whose masks keep hidden values out of the visible input; let the encoder attend along the dimensions of the operator output; and evaluate the view against its operator and a random encoder of the same architecture.

### 4.2 OpView: five views

We instantiate the method with five views, whose observation spaces hold the families of structure on which common EEG paradigms rely: signed morphology, spectral power, coupling between sensors,

polarity-free scalp states and short-term change (Appendix C). Each view below is given by its structure, observation space, removed nuisance and native target; $X \in \mathbb{R}^{C\times T}$ is a recording at 256 Hz with $C$ sensors. Each encoder (4.3 to 5.0 million parameters) attends along its operator's output dimensions.

**Waveform.** An event-related potential is a signed deflection, linear in the samples (Luck, 2014), so the Waveform operator is the identity on calibrated one-second patches and removes nothing; its observation space holds every linear function of the samples, such as a time-locked average. The target of a hidden patch $x_{c,n}$ of sensor $c$ is its spectrum under a Hann taper $w$, whitened by per-bin scales $\sigma$,

$$y_{c,n} = \operatorname{diag}(\sigma)^{-1}\mathcal{F}\big(w \odot x_{c,n}\big) \in \mathbb{C}^{129}, \tag{5}$$

which keeps the waveform information and spreads the loss evenly over frequency.

**Frequency.** A steady-state response is energy at physical frequencies (Lin et al., 2006), so the Frequency operator sets an observation space of power per frequency. With $F_{c,b}$ the orthonormal Fourier coefficient of sensor $c$ at bin $b$ over windows of one to four seconds, it exports, in 1 Hz groups,

$$P_{c,b} = \log\Big(1 + \frac{|F_{c,b}|^2}{s_b^2}\Big), \qquad r_b = \Big|\frac{1}{C}\sum_c \frac{F_{c,b}}{|F_{c,b}|}\Big|, \tag{6}$$

with scales $s_b$ fitted on training data, and removes each sensor's phase except its consistency $r_b$ across sensors. Every band power and harmonic sum is a simple function of $P$; its encoder reads it through a graph of neighbouring groups and harmonics, and predicts it for wholly hidden sensors.

**Relation.** Coupling between sensors (Nolte et al., 2004; Congedo et al., 2017) should not depend on the reference or the gain, so the Relation operator removes both. In a contrast basis $W$ ($W^\top \mathbf{1} = 0$) of $K$ smooth spatial functions around fixed anchors, the reference cancels, and the cross-spectral density $S_b$ of $W^\top X$ in each of six bands splits into a gain-free shape and a scale,

$$A_b = \log\big(S_b/q_b + 10^{-3} I\big), \qquad e_b = \log q_b, \qquad q_b = \operatorname{tr} S_b/K; \tag{7}$$

$A_b$ holds the coupling between every pair of anchor modes, and the encoder completes the rows of hidden anchor modes and predicts $e_b$ with a separate head.

**Projective.** Microstate analysis reads a scalp map regardless of its sign (Michel & Koenig, 2018), so the Projective operator removes the polarity. Each sample is charted in real spherical harmonics of degrees one to three, $a_t = B^+ x_t \in \mathbb{R}^{15}$, and each 16-sample token $I$ exports its log amplitude and

$$Q_I = \frac{\sum_{t\in I} \|a_t\|^2\, \vartheta_t \vartheta_t^\top}{\sum_{t\in I} \|a_t\|^2}, \qquad \vartheta_t = a_t/\|a_t\|, \tag{8}$$

which removes the sign, $Q_I(-a) = Q_I(a)$, the order of samples within the token and detail beyond degree three; the encoder predicts hidden states and amplitudes.

**Dynamics.** Some structure lies in how the signal is about to change, so the Dynamics operator sets an observation space of change at several time scales. It stacks the level $x_t$ with causal differences $\Delta_k x_t = x_t - x_{t-k}$, $k = 1, 2, 4, \ldots, 128$, and the encoder predicts $x_{t+h} - x_t$ for $h \in \{16, 32, 64\}$ samples, 62.5 to 250 ms ahead, from the past alone, the only target absent from its input.

## 5 Experiments

The experiments follow the argument of Sections 3 and 4: whether each encoder learns its pretraining task, whether this learning adds value beyond the operator and the architecture, how the views work as one system, and how they compare with published foundation models.

### 5.1 Setup

Each view is pretrained once, without labels, on a public EEG corpus for 40,000 or 160,000 updates (Appendix D). The corpus includes unlabelled recordings from seven evaluation datasets, marked

‡ as the benchmark marks such exposure. We evaluate on EEG-FM-Compass (Liu et al., 2026b) unchanged: nine datasets in six paradigms, four settings (frozen probing and fine-tuning, across subjects and within subject) and balanced accuracy, with differences below half a point counted as ties. A single view is read as the benchmark reads a published model (Appendix E); readouts over several views are fitted on training data only and reported separately (Appendix F). Aligned results (He & Wu, 2020) are compared only with EEGPT and MIRepNet, which the benchmark aligns.

### 5.2 Learning the pretraining tasks

Before measuring downstream benefit, we check that each encoder learned what its objective asks. On subjects of three corpus datasets who took no gradient step in pretraining, we compared each encoder's error on the hidden or future part with the best simple baseline that sees only the visible part, such as interpolation, the training mean or persistence (Figure 3a; Appendix D.1). Every encoder beats its best baseline on its main task, with median error ratios from 0.55 for forecasting 62.5 ms ahead, which zero-phase filtering may flatter, to 0.93 for completing masked scalp states. When the hidden part is large, such as a whole second, the encoders learn little beyond interpolation (Table 13).

### 5.3 Learning gain over the operator and the random encoder

Learning the pretraining task does not yet show that the representation is more useful than its operator. We therefore read three feature sets of a view under the same data, interface, readout and protocol: the operator output, the encoder with random weights and the pretrained encoder (Table 1). Pretrained minus random is the learning gain, and the margin over the operator is the stricter test. Pretraining improves on the random encoder in all ten fine-tuned comparisons and in six of eight frozen ones, and every one of these gains that has an operator score also exceeds the operator, by 1.1 to 19.4 points; on EEGMat, every pretrained view exceeds both controls frozen without alignment (Table 22).

Two views gain consistently. The Frequency encoder improves on its random counterpart in all eight of its comparisons (Table 1; Appendix H) and in all ten pairs read frozen by one ridge (Table 17), and the Dynamics encoder in all six fine-tuned ones, by 5.0 to 16.7 points, and frozen by up to 11.9; longer pretraining raises it further on eight of nine datasets (Appendix D).

The two exceptions are both frozen. On the P300 task BNCI2014009, the pretrained Waveform encoder trails its random counterpart by 6.2 points, mostly because of the probe: a longer probe without dropout narrows the gap to 1.9 points, and the P300 stays decodable (Appendix H.1). On BCI IV-2a, the Dynamics encoder trails by 1.0 point and gains 3.9 after 160,000 updates (Appendix D).

Table 1 uses the benchmark's readouts, which read random encoders poorly: on EEGMat, a closed-form ridge reads them up to 6.2 points better than the benchmark's probe, and pretrained ones about equally. Read by this ridge under a rule fixed in advance, on all 26 view–dataset pairs of six datasets frozen across subjects, pretraining still improves on the random encoder in 15 pairs without alignment and 18 with alignment: the Frequency encoder in all ten, the Dynamics encoder by up to 11.9 points, and the Waveform encoder in four of twelve (Table 17).

Where the gain is small or absent, two reasons fit our measurements. First, an objective strengthens only what it rewards (Proposition 1): completion and forecasting objectives favour structure predictable from context over transients locked to an external event, and the within-subject learning gain follows how much pretraining raises the share of feature variance that reflects the class, in 14 of 16 view–dataset pairs of a test fixed in advance ($\rho = 0.78$; Figure 9), an association (Appendix H). On P300, where the Waveform view's class share halves, objectives that reward such transients are the direct remedy. Second, where an operator already exposes a task's structure, little is left for learning to add: on EEGMat, the best operator alone exceeds the best published model frozen across subjects (Table 18), and under the ridge pretraining adds 0.6 to 3.9 points to each view there.

### 5.4 The view system at work

We check how the views divide the work and whether a view can be chosen without test labels. The views divide the work along their observation spaces. The two views whose operators remove nothing, Dynamics and Waveform, lead in 26 of the 36 dataset–setting cells, and each other view leads on tasks that match its observation space: Frequency on tasks defined by spectral power, Relation on

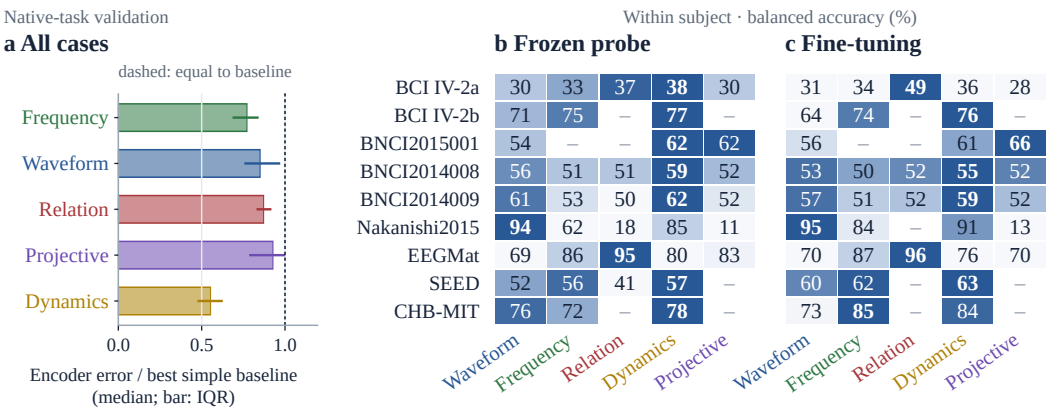



Figure 3: **Every encoder learns its pretraining task beyond simple baselines (a); within subject, the Dynamics view leads most often and the other views lead where their observation spaces hold the task's structure (b, c).** Bold: best view of each dataset.

Table 1: **Pretraining improves on the random encoder in every fine-tuned and six of eight frozen comparisons** (balanced accuracy, %). Op.: operator; Rand./Pre.: random/pretrained encoder.

| Dataset | View | Op. | Rand. | Pre. | Gain | Dataset | View | Op. | Rand. | Pre. | Gain |
|---|---|---|---|---|---|---|---|---|---|---|---|
| *LOSO, frozen* | | | | | | *LOSO, fine-tuned* | | | | | |
| EEGMat | Freq. | 71.7 | 68.2 | 73.1 | +4.9 | Nakanishi2015‡ | Dyn. | – | 74.3 | 91.0 | +16.7 |
| EEGMat | Rel. | 69.5 | 65.2 | 72.2 | +7.0 | Nakanishi2015‡ | Freq. | – | 70.7 | 83.9 | +13.1 |
| EEGMat | Dyn. | 66.9 | 59.0 | 68.0 | +9.0 | EEGMat | Dyn. | 67.1 | 59.1 | 70.1 | +11.0 |
| EEGMat | Wave. | 61.3 | 58.7 | 63.3 | +4.6 | BCI IV-2a | Dyn. | – | 37.6 | 46.0 | +8.4 |
| EEGMat | Proj. | 53.6 | 54.1 | 62.4 | +8.3 | BNCI2015001‡ | Dyn. | 57.2 | 58.4 | 63.4 | +5.0 |
| BCI IV-2a | Dyn. | – | 37.0 | 36.0 | −1.0 | *Within subject, fine-tuned* | | | | | |
| BNCI2014009‡ | Wave. | 69.2 | 69.2 | 63.0 | −6.2 | SEED‡ | Dyn. | 50.2 | 53.3 | 62.7 | +9.5 |
| *Within subject, probe* | | | | | | BCI IV-2b‡ | Dyn. | – | 64.1 | 76.5 | +12.4 |
| Nakanishi2015‡ | Wave. | – | 75.8 | 93.9 | +18.1 | EEGMat | Freq. | 67.7 | 76.3 | 87.1 | +10.8 |
| | | | | | | CHB-MIT‡ | Freq. | 70.4 | 77.2 | 84.6 | +7.4 |
| | | | | | | CHB-MIT‡ | Wave. | 58.9 | 67.5 | 72.8 | +5.3 |

tasks read from coupling between sensors, and Projective on one motor-imagery dataset (Table 2; Figure 3b,c; Appendix G). These leads are largest within subject, up to 20.0 points, and at most 3.5 points after fine-tuning across subjects: an observation space helps most where the learner has least to learn from, as Section 3 argued. On P300, a signed deflection (Luck, 2014), the three views whose operators discard the sign stay near chance, while the two that keep it lead.

Because the leading view changes with the task and the setting, the system needs a choice of view that uses no test label. Choosing for each subject the view whose frozen features score best on its calibration trials recovers about half of the gap between always using the Dynamics view and the best view in hindsight after fine-tuning (69.3 against 66.6 and 71.4), less under the probe (Appendix F.3).

### 5.5 Comparison with published foundation models

Read as the benchmark reads a published model, a single view beats the best of the twelve foundation models that EEG-FM-Compass evaluates in 10 of 36 dataset–setting cells without alignment, by up to 17.6 points, and the stronger of the two aligned models in 23 of 35 with alignment (Table 2).

The remaining gap lies mostly between subjects. Pretrained features give 1.2 to 6 times more variance to subject identity than random encoders do, which a cross-subject readout must ignore (Appendix H). Alignment removes this variance, lifting the frozen cross-subject multi-view results on emotion and seizure detection by 12.6 and 6.9 points; without alignment, four-class motor imagery trails in every setting.

Used together, the views match the strongest published model within half a point or exceed it on five of nine datasets frozen across subjects (Table 2, Fusion). These results remain comparable with

Table 2: **Comparison with the twelve foundation models of EEG-FM-Compass** (balanced accuracy, %). View: best single pretrained view under the benchmark's protocol; Fusion: views used without test labels, with operator and random-encoder features; FM: strongest published model (with alignment, EEGPT$^{E}$ or MIRepNet$^{M}$); Trad.: the benchmark's traditional decoder trained from scratch (Appendix E). Bold: ahead of FM by at least 0.5; other marks: Appendix E.

| | LOSO, frozen | | | LOSO, fine-tuned | | | | Within subject, probe | | | Within subject, fine-tuned | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *No alignment* | View | Fusion | FM | View | Fusion | FM | Trad. | View | Fusion | FM | View | Fusion | FM | Trad. |
| BCI IV-2a | 40.9$^{D}$ | 37.8$^{s}$ | 48.2 | 46.0$^{D}$ | 46.0 | 53.0 | 38.2 | 38.0$^{D}$ | 37.2 | 49.8 | 48.8$^{R}$ | 42.3 | 50.3 | 60.6 |
| BCI IV-2b‡ | **76.3**$^{D}$ | **77.5** | 75.6 | 75.6$^{D}$ | 75.6 | 77.7 | 73.2 | 76.9$^{D}$ | 76.8 | 76.7 | 76.5$^{D}$ | **77.9** | 77.4 | 79.2 |
| BNCI2015001‡ | **65.4**$^{D}$ | 63.4 | 63.4 | 63.4$^{D}$ | 63.4 | 64.1 | 56.4 | 61.6$^{D}$ | 60.7 | 67.8 | 65.5$^{P}$ | 63.8 | 70.3 | 75.9 |
| BNCI2014008‡ | 50.4$^{D}$ | 67.3 | 67.1 | 63.2$^{W}$ | 68.4$^{o}$ | 69.9 | 58.7 | 58.9$^{D}$ | 58.2 | 61.5 | 55.1$^{D}$ | 54.2 | 61.6 | 56.6 |
| BNCI2014009‡ | 69.6$^{D}$ | 69.7 | 74.1 | 71.1$^{W}$ | 77.1 | 77.3 | 65.1 | 62.5$^{D}$ | 62.1 | 65.8 | 59.0$^{D}$ | 59.2 | 65.5 | 54.5 |
| Nakanishi2015‡ | 80.6$^{F}$ | **87.7** | 85.5 | 91.0$^{D}$ | 93.3 | 92.9 | 94.1 | **93.9**$^{W}$ | **93.9** | 78.2 | **94.8**$^{W}$ | **94.8** | 85.9 | 98.8 |
| EEGMat | **72.6**$^{F}$ | **75.7** | 68.6 | **73.6**$^{F}$ | **73.6** | 72.6 | 67.4 | **94.7**$^{R}$ | **85.4** | 77.1 | **95.6**$^{R}$ | 86.1 | 86.1 | 95.6 |
| SEED‡ | 49.5$^{F}$ | 47.7 | 52.5‡ | 51.7$^{D}$ | 51.7 | 53.6 | 48.9 | **57.2**$^{D}$ | **55.9** | 52.3 | **62.7**$^{D}$ | **62.4** | 55.9 | 53.4 |
| CHB-MIT*‡ | 69.9$^{F}$ | 70.2 | 75.2 | 71.2$^{D}$ | 71.2 | 78.1 | 69.1 | 77.7$^{D}$ | 74.8 | 87.8 | 84.6$^{F}$ | 82.7 | 88.5 | 77.7 |
| *Alignment* | View | Fusion | FM | View | Fusion | FM | | View | Fusion | FM | View | Fusion | FM | |
| BCI IV-2a | 44.0$^{R}$ | **53.3** | 50.5$^{M}$ | **59.5**$^{D}$ | **60.0** | 54.2$^{M}$ | | 39.2$^{R}$ | 38.9 | 51.9$^{M}$ | *55.6*$^{D}$ | *45.3* | 63.3$^{M}$ | |
| BCI IV-2b‡ | 79.0$^{D}$ | 79.1 | 79.4$^{M}$ | ***82.3***$^{D}$ | ***82.3*** | 78.4$^{M}$ | | 76.9$^{D}$ | **77.7** | 76.8$^{M}$ | *80.4*$^{D}$ | *80.1* | 81.1$^{M}$ | |
| BNCI2015001‡ | 68.5$^{D}$ | 71.3 | 72.3$^{M}$ | ***78.0***$^{D}$ | ***78.0*** | 69.6$^{M}$ | | 66.7$^{P}$ | 64.9 | 73.4$^{M}$ | ***85.6***$^{D}$ | *73.2* | 80.8$^{M}$ | |
| BNCI2014008‡ | 50.4$^{D}$ | **68.3** | 57.3$^{E}$ | **68.5**$^{W}$ | **68.5** | 58.2$^{E}$ | | **58.2**$^{D}$ | **56.6** | 51.3$^{E}$ | ***54.1***$^{D}$ | ***53.1*** | 51.9$^{E}$ | |
| BNCI2014009‡ | **75.5**$^{D}$ | **76.9** | 66.5$^{E}$ | ***74.2***$^{W}$ | ***74.2*** | 62.8$^{E}$ | | **58.6**$^{D}$ | **58.0** | 53.2$^{E}$ | *52.2*$^{D}$ | *51.9* | 52.0$^{E}$ | |
| Nakanishi2015‡ | 75.1$^{F}$ | **88.6** | 85.5$^{E}$ | **93.1**$^{D}$ | **93.1** | 88.3$^{E}$ | | **93.3**$^{W}$ | **91.9** | 77.8$^{E}$ | ***100.0***$^{D}$ | ***100.0*** | 75.7$^{E}$ | |
| EEGMat | **82.3**$^{F}$ | **86.4** | 59.4$^{E}$ | ***70.5***$^{F}$ | ***70.5*** | 58.0$^{E}$ | | **94.8**$^{R}$ | **86.7** | 67.1$^{E}$ | ***100.0***$^{R}$ | ***97.6*** | 65.6$^{E}$ | |
| SEED‡ | **58.1**$^{F}$ | **60.3** | 50.0‡$^{E}$ | *47.2*$^{D}$ | *47.2* | 48.7‡$^{E}$ | | **64.5**$^{F}$ | **64.5** | 42.9‡$^{E}$ | – | – | 40.5‡$^{E}$ | |
| CHB-MIT*‡ | **73.5**$^{D}$ | **77.1** | 70.9$^{E}$ | ***74.1***$^{F}$ | ***74.1*** | 66.9$^{E}$ | | **70.9**$^{D}$ | **72.6** | 66.7$^{E}$ | ***82.2***$^{F}$ | ***77.8*** | 63.7$^{E}$ | |

the published ones: the fusion readouts keep the benchmark's data, splits and metric, use no test label and read the views with linear classifiers, like the benchmark's probe, or with combinations of them, while the FM column takes the best of twelve models in each cell in hindsight (Appendix E). Since they also draw on operator and random-encoder features, they measure the whole system; they gain most where the benchmark's probe reads single views poorly, as on BNCI2014008, where the Waveform operator alone beats the combination, so this match needs no pretraining (Appendix F).

## 6 Conclusion

Our results suggest that the feature space in which an EEG foundation model learns deserves the same attention as its data, architecture and objective. A fixed operator reads a known power change more accurately than the frontends and pretrained encoders of released models, and changing only the feature space or only the objective changes what the same budget learns. Operator preconditioning acts on both: in OpView, every encoder learns its native task, the views divide the work by observation space, single views are competitive with published foundation models, pretraining improves every fine-tuned comparison by 5 to 17 points and pays off consistently for the Frequency and Dynamics views; where it adds little, the objective does not yet reward what the task needs, or the operator already exposes it. More broadly, the approach turns what EEG research has learned to measure into a design space for foundation models: each measurable structure, from sleep spindles to cortical sources, can become a view with its own observation space, invariances and objective, added without retraining the others and held to account against its operator and a random encoder, so that model capacity goes to the structures EEG decoding uses.

**Limitations.** Two questions remain open. First, an objective defined in a view's native space need not reward the structure a task needs. Targets, masks and time scales should be designed together for each view, especially for transient responses that prediction from context underweights, and judged against the operator and a matched random encoder, not by the pretraining loss alone. Second, OpView pretrains each view on its own and learns to combine them only on the downstream task. A fusion module pretrained on unlabelled EEG could learn how the views interact and give downstream tasks one compact input. It would have to keep what each view captures alone, not only what the views share, accept any subset of views, and still let one encoder be added or replaced without retraining the others.

## AI USE STATEMENT

Large language models were used throughout this work. Coding agents based on Claude (Anthropic) and GPT (OpenAI) models implemented the operators, encoders and evaluation code and ran the experiments under the authors' direction, and Claude drafted and revised the text. The authors set the research questions, made the design decisions and take full responsibility for the content.

## ETHICS STATEMENT

This work uses only publicly released EEG recordings; we collected no new data and ran no experiments with participants. Each dataset was released by its original authors for research, and we used the recordings under their licence and data-use terms (Appendix D), without attempting to identify any participant. Models that decode EEG can reveal information about a person's state, so applying them to new recordings should require that person's informed consent.

## REPRODUCIBILITY STATEMENT

Appendix C specifies every operator, encoder, target, mask and auxiliary term; Appendix D the pretraining data and optimisation; Appendix E the evaluation protocol and every departure from EEG-FM-Compass; Appendix F the readouts that use several views; and Appendix G every per-view result.

## REFERENCES

G. Alain and Y. Bengio. Understanding Intermediate Layers Using Linear Classifier Probes. ICLR Workshop, 2017.

S. Amari. Natural Gradient Works Efficiently in Learning. Neural Computation, 10(2):251–276, 1998.

K. K. Ang et al. Filter Bank Common Spatial Pattern Algorithm on BCI Competition IV Datasets 2a and 2b. Frontiers in Neuroscience, 6:39, 2012. URL https://doi.org/10.3389/fnins.2012.00039.

S. Arora, S. S. Du, W. Hu, Z. Li, and R. Wang. Fine-Grained Analysis of Optimization and Generalization for Overparameterized Two-Layer Neural Networks. ICML, 2019.

R. Balestriero and Y. LeCun. How Learning by Reconstruction Produces Uninformative Features for Perception. ICML, 2024.

H. Banville, O. Chehab, A. Hyvärinen, D.-A. Engemann, and A. Gramfort. Uncovering the Structure of Clinical EEG Signals with Self-Supervised Learning. Journal of Neural Engineering, 18(4):046020, 2021.

A. Barachant, S. Bonnet, M. Congedo, and C. Jutten. Multiclass Brain–Computer Interface Classification by Riemannian Geometry. IEEE Transactions on Biomedical Engineering, 59(4):920–928, 2012.

J. S. Bindra and S. Panwar. A spectral audit framework reveals task-dependent aperiodic reliance across EEG and ECG deep learning, 2026. URL https://arxiv.org/abs/2606.08583.

C. M. Bishop. Pattern Recognition and Machine Learning. Springer, 2006.

B. Blankertz, R. Tomioka, S. Lemm, M. Kawanabe, and K.-R. Müller. Optimizing Spatial Filters for Robust EEG Single-Trial Analysis. IEEE Signal Processing Magazine, 25(1):41–56, 2008.

M. M. Bronstein, J. Bruna, T. Cohen, and P. Veličković. Geometric Deep Learning: Grids, Groups, Graphs, Geodesics, and Gauges. arXiv:2104.13478, 2021.

T. Cohen and M. Welling. Group Equivariant Convolutional Networks. ICML, 2016.

M. Congedo, A. Barachant, and R. Bhatia. Riemannian Geometry for EEG-based Brain–Computer Interfaces; a Primer and a Review. Brain–Computer Interfaces, 4(3):155–174, 2017.

T. M. Cover and J. A. Thomas. Elements of Information Theory. 2nd edition, Wiley, 2006.

W. Cui, W. Jeong, P. Thölke, T. Medani, K. Jerbi, A. A. Joshi, and R. M. Leahy. Neuro-GPT: Towards a Foundation Model for EEG. IEEE ISBI, 2024.

J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. NAACL, 2019.

B. Döner, T. M. Ingolfsson, L. Benini, and Y. Li. LUNA: Efficient and Topology-Agnostic Foundation Model for EEG Signal Analysis. Advances in Neural Information Processing Systems, 2025. URL https://arxiv.org/abs/2510.22257.

T. Donoghue et al. Parameterizing neural power spectra into periodic and aperiodic components. Nature Neuroscience, 2020. URL https://doi.org/10.1038/s41593-020-00744-x.

M. Federici, A. Dutta, P. Forré, N. Kushman, and Z. Akata. Learning Robust Representations via Multi-View Information Bottleneck. ICLR, 2020.

R. Geirhos et al. Shortcut Learning in Deep Neural Networks. Nature Machine Intelligence, 2:665–673, 2020.

A. L. Goldberger et al. PhysioBank, PhysioToolkit, and PhysioNet: Components of a New Research Resource for Complex Physiologic Signals. Circulation, 101(23):e215–e220, 2000.

H. He and D. Wu. Transfer Learning for Brain–Computer Interfaces: A Euclidean Space Data Alignment Approach. IEEE Transactions on Biomedical Engineering, 67(2):399–410, 2020.

K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick. Masked Autoencoders Are Scalable Vision Learners. CVPR, 2022.

J. Hewitt and P. Liang. Designing and Interpreting Probes with Control Tasks. EMNLP, 2019.

S. Ioffe and C. Szegedy. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. ICML, 2015.

A. Jacot, F. Gabriel, and C. Hongler. Neural Tangent Kernel: Convergence and Generalization in Neural Networks. NeurIPS, 2018. URL https://arxiv.org/abs/1806.07572.

W.-B. Jiang, L.-M. Zhao, and B.-L. Lu. Large Brain Model for Learning Generic Representations with Tremendous EEG Data in BCI. ICLR, 2024. URL https://openreview.net/forum?id=QzTpTRVtrP.

D. P. Kingma and J. Ba. Adam: A Method for Stochastic Optimization. ICLR, 2015.

D. Kostas, S. Aroca-Ouellette, and F. Rudzicz. BENDR: Using Transformers and a Contrastive Self-Supervised Learning Task to Learn from Massive Amounts of EEG Data. Frontiers in Human Neuroscience, 15, 2021.

A. Kumar, A. Raghunathan, R. Jones, T. Ma, and P. Liang. Fine-Tuning Can Distort Pretrained Features and Underperform Out-of-Distribution. ICLR, 2022a.

V. Kumar, L. Reddy, S. K. Sharma, K. Dadi, C. Yarra, B. S. Raju, and S. Rajendran. mulEEG: A Multi-View Representation Learning on EEG Signals, 2022b. URL https://arxiv.org/abs/2204.03272.

V. J. Lawhern et al. EEGNet: A Compact Convolutional Neural Network for EEG-based Brain–Computer Interfaces. Journal of Neural Engineering, 15(5):056013, 2018.

Y. LeCun, L. Bottou, G. B. Orr, and K.-R. Müller. Efficient BackProp. In Neural Networks: Tricks of the Trade, Springer, 1998.

J. D. Lee, Q. Lei, N. Saunshi, and J. Zhuo. Predicting What You Already Know Helps: Provable Self-Supervised Learning. NeurIPS, 2021.

Z. Lin, C. Zhang, W. Wu, and X. Gao. Frequency Recognition Based on Canonical Correlation Analysis for SSVEP-based BCIs. IEEE Transactions on Biomedical Engineering, 53(12):2610–2614, 2006.

D. Liu, Z. Chen, J. Luo, et al. MIRepNet: A Pipeline and Pre-trained Model for EEG-Based Motor Imagery Classification. Knowledge-Based Systems, 343:115966, 2026a.

D. Liu et al. EEG-FM-Compass: Progress, Benchmarking, and Future Directions for EEG Foundation Models, 2026b. URL https://arxiv.org/abs/2601.17883.

F. Lotte et al. A Review of Classification Algorithms for EEG-based Brain–Computer Interfaces: A 10 Year Update. Journal of Neural Engineering, 15(3):031005, 2018.

S. J. Luck. An Introduction to the Event-Related Potential Technique. 2nd edition, MIT Press, 2014.

J. Martens and R. Grosse. Optimizing Neural Networks with Kronecker-factored Approximate Curvature. ICML, 2015.

C. M. Michel and T. Koenig. EEG Microstates as a Tool for Studying the Temporal Dynamics of Whole-Brain Neuronal Networks: A Review. NeuroImage, 180:577–593, 2018.

M. Nakanishi, Y. Wang, Y.-T. Wang, and T.-P. Jung. A Comparison Study of Canonical Correlation Analysis Based Methods for Detecting Steady-State Visual Evoked Potentials. PLoS ONE, 10(10):e0140703, 2015.

M. Nakanishi, Y. Wang, X. Chen, Y.-T. Wang, X. Gao, and T.-P. Jung. Enhancing Detection of SSVEPs for a High-Speed Brain Speller Using Task-Related Component Analysis. IEEE Transactions on Biomedical Engineering, 65(1):104–112, 2018.

J. Nocedal and S. J. Wright. Numerical Optimization. 2nd edition, Springer, 2006.

G. Nolte et al. Identifying True Brain Interaction from EEG Data Using the Imaginary Part of Coherency. Clinical Neurophysiology, 115(10):2292–2307, 2004.

M. Pezeshki et al. Gradient Starvation: A Learning Proclivity in Neural Networks. NeurIPS, 2021. URL https://arxiv.org/abs/2011.09468.

G. Pfurtscheller and F. H. Lopes da Silva. Event-related EEG/MEG synchronization and desynchronization: basic principles. Clinical Neurophysiology, 110(11):1842–1857, 1999. URL https://doi.org/10.1016/S1388-2457(99)00141-8.

N. Rahaman et al. On the Spectral Bias of Neural Networks. ICML, 2019.

A. Rahimi and B. Recht. Random Features for Large-Scale Kernel Machines. NeurIPS, 2007.

Y. Roy et al. Deep Learning-based Electroencephalography Analysis: A Systematic Review. Journal of Neural Engineering, 16(5):051001, 2019.

A. M. Saxe, P. W. Koh, Z. Chen, M. Bhand, B. Suresh, and A. Y. Ng. On Random Weights and Unsupervised Feature Learning. ICML, 2011.

R. T. Schirrmeister et al. Deep Learning with Convolutional Neural Networks for EEG Decoding and Visualization. Human Brain Mapping, 38(11):5391–5420, 2017.

B. Schölkopf and A. J. Smola. Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond. MIT Press, 2002.

H. Shah, K. Tamuly, A. Raghunathan, P. Jain, and P. Netrapalli. The Pitfalls of Simplicity Bias in Neural Networks. NeurIPS, 2020.

A. H. Shoeb. Application of Machine Learning to Epileptic Seizure Onset Detection and Treatment. PhD thesis, Massachusetts Institute of Technology, 2009.

M. Tancik et al. Fourier Features Let Networks Learn High Frequency Functions in Low Dimensional Domains. NeurIPS, 2020. URL https://arxiv.org/abs/2006.10739.

L. Tang et al. What Do EEG Foundation Models Capture from Human Brain Signals? arXiv:2605.11410, 2026. URL https://arxiv.org/abs/2605.11410.

M. Tangermann et al. Review of the BCI Competition IV. Frontiers in Neuroscience, 6:55, 2012.

Y. Tian, D. Krishnan, and P. Isola. Contrastive Multiview Coding. ECCV, 2020.

Y.-H. H. Tsai, Y. Wu, R. Salakhutdinov, and L.-P. Morency. Self-supervised Learning from a Multi-view Perspective. ICLR, 2021.

G. Wang et al. EEGPT: Pretrained Transformer for Universal and Reliable Representation of EEG Signals. NeurIPS, 2024. URL https://papers.nips.cc/paper_files/paper/2024/hash/4540d267eeec4e5dbd9dae9448f0b739-Abstract-Conference.html.

J. Wang et al. CBraMod: A Criss-Cross Brain Foundation Model for EEG Decoding. ICLR, 2025. URL https://arxiv.org/abs/2412.07236.

P. Welch. The Use of Fast Fourier Transform for the Estimation of Power Spectra. IEEE Transactions on Audio and Electroacoustics, 15(2):70–73, 1967.

W. Xiong et al. EEG-FM-Bench: A Comprehensive Benchmark for the Systematic Evaluation and Diagnostic Analyses of EEG Foundation Models, 2025. URL https://arxiv.org/abs/2508.17742.

Y. Xu, S. Zhao, J. Song, R. Stewart, and S. Ermon. A Theory of Usable Information under Computational Constraints. ICLR, 2020.

C. Yang, M. B. Westover, and J. Sun. BIOT: Biosignal Transformer for Cross-data Learning in the Wild. NeurIPS, 2023.

L. Yang, Q. Sun, A. Li, and M. M. Van Hulle. Are EEG Foundation Models Worth It? Comparative Evaluation with Traditional Decoders in Diverse BCI Tasks. International Conference on Learning Representations, 2026. URL https://openreview.net/forum?id=5Xwm8e6vbh.

W.-L. Zheng and B.-L. Lu. Investigating Critical Frequency Bands and Channels for EEG-based Emotion Recognition with Deep Neural Networks. IEEE Transactions on Autonomous Mental Development, 7(3):162–175, 2015.

I. Zyma et al. Electroencephalograms during Mental Arithmetic Task Performance. Data, 4(1):14, 2019.

# A PROOFS

This appendix proves the statements used in Sections 3 and 4. Throughout, $U$ is the visible context of a prediction problem, $T \in \mathbb{R}^d$ its target with finite second moment, $Q = q(U)$ a simpler statistic of the context and $A = a(U)$ a structural statistic, and $\|h\|_2^2 = \mathbb{E}\|h(U)\|^2$. The statements concern the population squared loss and explain the direction of effects; the scores of finite classifiers are measured in the experiments.

## A.1 LINEAR READOUT OF AN EVEN STRUCTURE

Section 3.1 uses band power to show that the feature space alone can decide readability. Suppose $X \overset{d}{=} -X$, $a(X) = a(-X)$, and $X$ and $a(X)$ have finite second moments. Symmetry gives $\mathbb{E}X = 0$ and $\mathbb{E}[Xa(X)] = -\mathbb{E}[Xa(X)] = 0$, so for any affine predictor $w^\top X + c$,

$$\mathbb{E}\big(a(X) - w^\top X - c\big)^2 = \mathrm{Var}\left(a(X)\right) + \big(c - \mathbb{E}a(X)\big)^2 + \mathbb{E}(w^\top X)^2,$$

which the constant $\mathbb{E}a(X)$ minimises. If $a$ is affine in $O(X)$, an affine readout of $O(X)$ recovers it exactly; the band power $\|Bx\|^2$ is such a structure for the operator $O(x) = \|Bx\|^2$. The statement concerns the readout class; a nonlinear encoder can compute the power, and whether pretraining leads it to do so is the subject of Section 3.2.

### A.2 Proposition 1: reward of a squared-loss objective

For $V = Q$ or $V = (Q, A)$, orthogonal projection in $L^2$ gives $\mathbb{E}\|T - g(V)\|^2 = \mathbb{E}\|T - \mathbb{E}[T \mid V]\|^2 + \mathbb{E}\|\mathbb{E}[T \mid V] - g(V)\|^2$, so the optimal expected loss given $V$ is $\mathbb{E}\|T - \mathbb{E}[T \mid V]\|^2$. The two information sets are nested, and subtracting the two optimal losses gives $\mathbb{E}\|m_{QA} - m_Q\|^2 = \mathbb{E}\|\delta\|^2$.

The reward also sets the gradient. If a predictor $f_\theta = m_Q$ is differentiable along a direction $v$ in $L^2$ and $h_v = D_v f_\theta$ is measurable in $(Q, A)$, the directional derivative of $L = \frac{1}{2}\mathbb{E}\|f_\theta - T\|^2$ is $\mathbb{E}\langle h_v, m_Q - T\rangle$; because $h_v$ is measurable in $(Q, A)$, $T$ can be replaced by $m_{QA}$, which gives $D_v L = -\mathbb{E}\langle h_v, \delta\rangle$. Since $\mathbb{E}[\delta \mid Q] = 0$, only the part of $h_v$ not explained by $Q$ contributes, and the Cauchy–Schwarz inequality gives $|D_v L| \le \left(\mathbb{E}\|h_v - \mathbb{E}[h_v \mid Q]\|^2\, \Delta_T(A \mid Q)\right)^{1/2}$. □

Both conditions of this gradient statement are needed, as two examples show. With independent symmetric signs $A, W, U = (A, W)$, constant $Q$ and $T = W$, the reward of $A$ is zero, yet $f_\theta = \theta W$ has derivative $-1$ at zero, because the direction uses context outside $(Q, A)$. Conversely, $T = A$ and $f_\theta = \theta^2 A$ give a positive reward and a zero derivative at initialisation. The reward measures what the population offers, and the derivative measures what an optimiser can reach.

**A waveform target does not reward predictable power.** The first consequence in Section 3.2 follows from a simple model of an ongoing rhythm. Suppose a latent state $Z$ controls a future segment $X'$ with $\mathbb{E}[X' \mid U, Z] = 0$ and $\mathbb{E}[X'X'^{\top} \mid U, Z] = \Sigma(Z)$. The optimal waveform prediction is $\mathbb{E}[X' \mid U] = 0$ whatever the context, so a waveform target offers no reward for any statistic of $U$. A second-order target $X'X'^{\top}$ has optimal prediction $\mathbb{E}[\Sigma(Z) \mid U]$, and its reward for $A$ beyond $Q$ is $\mathbb{E}\|\mathbb{E}[\Sigma(Z) \mid Q, A] - \mathbb{E}[\Sigma(Z) \mid Q]\|_F^2$, positive whenever the context predicts the power. This describes a rhythm whose phase is unpredictable and whose power is predictable.

The ablation of Appendix B.3 varies this reward directly. Multiplying the Fourier vector $C_k$ of a target by an independent unit-modulus $\eta_k$ with mean $\kappa$, shared across channels, gives $\Delta_{\eta C}(A \mid Q) = |\kappa|^2 \Delta_C(A \mid Q)$ and leaves power and same-frequency cross-spectra unchanged. For a whole-patch polarity $\eta \in \{-1, 1\}$ the sign-averaged loss is

$$\mathbb{E}_\eta\|f - \eta x\|^2 = \|f - \kappa x\|^2 + (1 - \kappa^2)\|x\|^2, \tag{9}$$

so at $\kappa = 0$ the waveform target only shrinks its prediction, while a second-order target keeps rewarding predictable structure.

### A.3 Reward a model can collect, and learning speed

The second consequence in Section 3.2 concerns how much of the reward a model can collect. Let $\mathcal{S} \subset L^2(Q, A; \mathbb{R}^d)$ be a closed subspace of realisable directions at $f = m_Q$, so that the linearised model class is $m_Q + \mathcal{S}$, and let $R(f) = \mathbb{E}\|T - f\|^2$ and $P_\mathcal{S}$ the orthogonal projection onto $\mathcal{S}$. For $h \in \mathcal{S}$, the residual $T - m_{QA}$ is orthogonal to $h$ and to $\delta$, so $R(m_Q + h) = \mathbb{E}\|T - m_{QA}\|^2 + \|\delta - h\|_2^2$. The minimiser over the closed subspace is $h = P_\mathcal{S}\delta$, and $R(m_Q) - \inf_{h \in \mathcal{S}} R(m_Q + h) = \|P_\mathcal{S}\delta\|_2^2$; the projection is a contraction, so the collected share $\rho_\mathcal{S} = \|P_\mathcal{S}\delta\|_2^2 / \|\delta\|_2^2$ lies in $[0, 1]$. Finally $-D_h L(m_Q) = \langle h, \delta\rangle$, whose supremum over the unit ball of $\mathcal{S}$ is $\|P_\mathcal{S}\delta\|_2$. □

The target thus decides $\Delta_T$, and the feature space and the model decide $\mathcal{S}$ and therefore $\rho_\mathcal{S}$. The feature space also sets the speed of learning. Gradient descent on squared loss in a fixed-kernel model from zero gives $f^{(t)} - y = -(I - \eta K)^t y$. In the eigenbasis of $K$, with step $\eta \le 1/\lambda_{\max}$, the error along eigendirection $u_j$ decays as

$$\|f^{(t)} - y\|^2 = \sum_j c_j^2 (1 - \eta\lambda_j)^{2t}, \qquad y = \sum_j c_j u_j, \tag{10}$$

so the part of a target on small-eigenvalue directions can stay unlearned within a finite budget, and an input representation that raises those eigenvalues speeds learning with the target unchanged. Figure 4A shows this for a two-mode kernel under exact gradient descent with step 0.5, where the slow mode's eigenvalue is raised from 0.02 to 1. Figure 4B repeats the comparison in a constructed EEG example: predictors of 13,776 parameters learn, for 150 updates and five seeds, a known interaction between adjacent cortical sources under an fsaverage forward model, given either the interaction itself or, as controls, raw inputs, a product of non-adjacent sources or an adjacency with

permuted geometry. When two input paths can explain the same target direction, gradient descent splits the explained part between them in the ratio of their eigenvalues; giving each view its own objective avoids this competition.

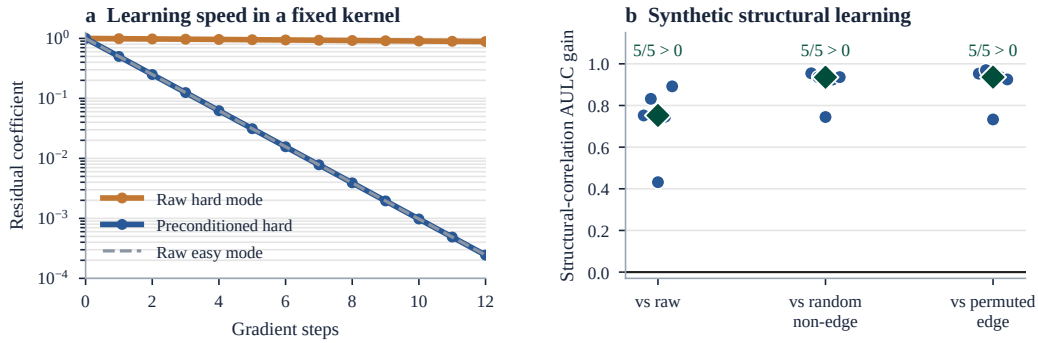



Figure 4: **The feature space sets how fast a fixed target is learned.** Raising an eigenvalue speeds learning of its direction (A), and making a known cortical interaction explicit speeds its learning (B).

### A.4 Invariance

Section 4.1 argues that every invariance of an operator has a price. If $O(gX) = O(X)$ for a nuisance action $g$, every head on $O$ inherits the invariance: its expected loss is unchanged across environments that differ only by such actions, and labels that differ only along an orbit of $g$ receive the same prediction. For a contrast basis with $W^\top \mathbf{1} = 0$, a common reference disappears from $W^\top X$; a common gain multiplies the cross-spectral density and its trace by the same factor, so the shape $A_b$ of Equation 7 is gain invariant and the scale $e_b$ carries exactly what the invariance removes. The raw band power satisfies $|F_{c,b}(gX)|^2 = g^2|F_{c,b}(X)|^2$, so a contrast of $\log|F_{c,b}|^2$ whose weights sum to zero is exactly invariant to a common gain. For the floored power of Equation 6, $P_{c,b}(gX) = \log[1 + g^2(e^{P_{c,b}(X)} - 1)]$, and the invariance holds approximately, where $|F_{c,b}|^2 \gg s_b^2$.

### A.5 A joint head and the best single view

Appendix F combines views, and this bound states what a combination can achieve. Let $\mathcal{R}_p(Z)$ be the lowest expected loss on task $p$ of a declared head class on $Z$, and let the joint head class on $\mathcal{Z} = [Z_v]_{v\in\mathcal{V}}$ contain every single-view head. For task weights $\pi_p \geq 0$,

$$\sum_p \pi_p \mathcal{R}_p(\mathcal{Z}) \leq \sum_p \pi_p \min_v \mathcal{R}_p(Z_v) \leq \min_v \sum_p \pi_p \mathcal{R}_p(Z_v).$$

The first inequality holds because each single-view head is available in the joint class, the second because the minimum over views can be taken per task; it is strict whenever no single view attains the minimum on every weighted task. A fitted head adds its own excess loss over its class optimum, so the inequality bounds what a readout can reach on infinite data; on finite data the fitting rule matters, which is why the rules of Appendix F are fixed in advance.

## B Released models and controlled studies

This appendix gives the setup, full results and limits of each experiment in Section 3, in the order of the main text.

### B.1 Patch embeddings and complete encoders of released models

This analysis asks whether pretraining keeps the task information that a model's own patch embeddings carry (Section 3.3). For five released models (EEGPT, CBraMod, LaBraM-Base, LUNA-Base, ST-EEGFormer-Base) we compared the learned patch embeddings $F$ before the first attention block with the complete frozen encoder $H$, each pretrained and at three random initialisations; the

patch embeddings are themselves learned. The datasets are BCI IV-2a (5,184 trials, both sessions), BNCI2015001 (5,600), Nakanishi2015 (1,620) and ERP CORE P3 (8,000), 20 model–dataset pairs in all. Readouts are closed-form ridge, with penalties chosen on three inner folds inside five subject-disjoint outer folds; means, scales and projections are fitted on training splits, and model weights are frozen. Each model is read through its own input pipeline (sampling rate, scaling and patching), so the analysis compares modules within each model, with folds and cohorts of its own.

Table 3: **Depth gap and pretraining gain of each released model.** $H - F$: encoder minus patch embeddings; gain: pretrained minus random. Bold: 95% interval excludes zero; *seen in pretraining.

| | BCI IV-2a | | BNCI2015001 | | Nakanishi2015 | | ERP CORE P3 | |
|---|---|---|---|---|---|---|---|---|
| Model | $H - F$ | gain | $H - F$ | gain | $H - F$ | gain | $H - F$ | gain |
| EEGPT | **−6.7** | −0.9 | **−3.6** | **+2.6** | **−16.0** | +1.4 | **−4.7** | **+1.8** |
| CBraMod | −1.4 | **+8.7** | +1.2 | **+5.4** | **−21.9** | **+14.3** | **−3.2** | +1.3 |
| LaBraM | +0.8 | **+2.3** | +1.5 | **+2.5** | **−21.0** | −2.8 | **−1.9** | **−2.5** |
| LUNA | **−4.8** | **+3.6** | **−3.7** | **+3.0** | **−21.4** | **−17.5** | **−5.7** | **+5.2** |
| ST-EEGFormer | **+20.6*** | **+14.4*** | **+10.3** | **+7.2** | **+12.0** | **+15.0** | **+7.6** | **+1.5** |

Table 4: **Four of five pretrained encoders fall 16–22 points below their own patch embeddings on steady-state responses.** Subject-bootstrap intervals; Holm $p$ over 24 planned contrasts. *ST-EEGFormer saw BCI IV-2a in pretraining.

| Model | $F_{pre}$ | $H_{pre}$ | $H - F$ [95% interval] | Holm $p$ |
|---|---|---|---|---|
| BCI IV-2a: MI (BCA; 9 subjects; 5184 trials) | | | | |
| EEGPT | 35.76 | 29.11 | -6.66 [-8.51, -4.78] | 0.07812 |
| CBraMod | 43.79 | 42.44 | -1.35 [-3.05, 0.21] | 1.00000 |
| LaBraM | 36.07 | 36.92 | +0.85 [-2.08, 3.84] | 1.00000 |
| LUNA | 38.58 | 33.76 | -4.82 [-6.54, -3.40] | 0.07812 |
| ST-EEGFormer* | 33.68 | 54.24 | +20.56 [11.40, 28.99] | 0.10986 |
| BNCI2015001: MI (BCA; 12 subjects; 5600 trials) | | | | |
| EEGPT | 60.26 | 56.71 | -3.56 [-5.25, -1.63] | 0.12891 |
| CBraMod | 61.22 | 62.46 | +1.24 [-0.89, 3.35] | 1.00000 |
| LaBraM | 59.12 | 60.60 | +1.48 [-1.05, 4.61] | 1.00000 |
| LUNA | 61.79 | 58.14 | -3.65 [-5.90, -1.31] | 0.15625 |
| ST-EEGFormer | 58.52 | 68.83 | +10.31 [4.59, 16.25] | 0.10986 |
| Nakanishi2015: SSVEP (BCA; 9 subjects; 1620 trials) | | | | |
| EEGPT | 87.96 | 71.98 | -15.99 [-26.98, -6.23] | 0.12891 |
| CBraMod | 77.53 | 55.62 | -21.91 [-29.57, -14.63] | 0.07812 |
| LaBraM | 80.68 | 59.69 | -20.99 [-26.79, -15.74] | 0.07812 |
| LUNA | 51.73 | 30.37 | -21.36 [-29.07, -13.77] | 0.07812 |
| ST-EEGFormer | 79.20 | 91.17 | +11.98 [5.68, 19.26] | 0.10986 |
| ERP CORE P3: ERP (AUROC; 40 subjects; 8000 trials) | | | | |
| EEGPT | 73.00 | 68.25 | -4.74 [-6.46, -3.09] | 0.00024 |
| CBraMod | 76.57 | 73.35 | -3.21 [-4.37, -2.04] | 0.00063 |
| LaBraM | 75.27 | 73.40 | -1.87 [-3.49, -0.26] | 0.27990 |
| LUNA | 77.22 | 71.52 | -5.70 [-6.85, -4.56] | 0.00024 |
| ST-EEGFormer | 71.56 | 79.18 | +7.62 [6.08, 9.17] | 0.00024 |

Only on ERP CORE P3, with 40 subjects, do contrasts pass a Holm correction over the 24 planned contrasts: four of the five, three of them losses; the other cohorts have nine to twelve subjects. On steady-state responses four of the five encoders lose 16 to 22 points to their patch embeddings, while ST-EEGFormer gains 12. Relative to random initialisation the complete encoder gains in 16 of 20 pairs (Table 5). The depth gap $H - F$ is negative in 13 pairs when pretrained and in 15 at random initialisation; pretraining narrows it, $(H - F)_{\mathrm{pre}} - (H - F)_{\mathrm{rand}} > 0$, in 14 pairs and widens it in 6: EEGPT, CBraMod, LaBraM and LUNA on Nakanishi2015 (by 3.7, 4.3, 23.7 and 17.6 points) and CBraMod and LaBraM on ERP CORE P3 (0.7 and 2.4). Figure 5 shows the subject-level differences, and Table 6 the secondary endpoints at fixed rank, under which the encoder falls below its patch embeddings in 9, 7 and 11 of the 20 pairs at ranks 16, 64 and 256. These counts describe point estimates.

Table 5: **Patch embeddings and encoders, pretrained and random** (mean of three seeds), balanced accuracy or AUROC (%).

| Dataset | Model | $F$, pretrained | $F$, random | $H$, pretrained | $H$, random |
|---|---|---|---|---|---|
| BCI IV-2a | EEGPT | 35.76 | 39.83 | 29.11 | 29.97 |
| | CBraMod | 43.79 | 35.71 | 42.44 | 33.69 |
| | LaBraM | 36.07 | 34.27 | 36.92 | 34.65 |
| | LUNA | 38.58 | 35.06 | 33.76 | 30.13 |
| | ST-EEGFormer* | 33.68 | 40.59 | 54.24 | 39.83 |
| BNCI2015001 | EEGPT | 60.26 | 60.67 | 56.71 | 54.08 |
| | CBraMod | 61.22 | 59.36 | 62.46 | 57.03 |
| | LaBraM | 59.12 | 58.82 | 60.60 | 58.05 |
| | LUNA | 61.79 | 60.34 | 58.14 | 55.13 |
| | ST-EEGFormer | 58.52 | 61.17 | 68.83 | 61.59 |
| Nakanishi2015 | EEGPT | 87.96 | 82.92 | 71.98 | 70.62 |
| | CBraMod | 77.53 | 58.87 | 55.62 | 41.28 |
| | LaBraM | 80.68 | 59.73 | 59.69 | 62.47 |
| | LUNA | 51.73 | 51.63 | 30.37 | 47.84 |
| | ST-EEGFormer | 79.20 | 81.50 | 91.17 | 76.15 |
| ERP CORE P3 | EEGPT | 73.00 | 76.42 | 68.25 | 66.46 |
| | CBraMod | 76.57 | 74.57 | 73.35 | 72.09 |
| | LaBraM | 75.27 | 75.35 | 73.40 | 75.89 |
| | LUNA | 77.22 | 74.77 | 71.52 | 66.33 |
| | ST-EEGFormer | 71.56 | 76.62 | 79.18 | 77.68 |

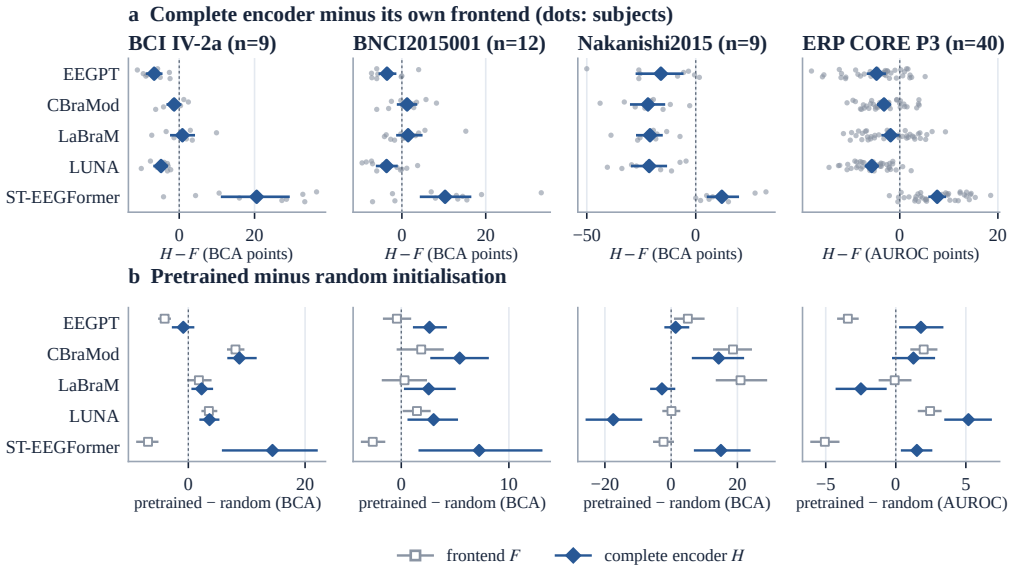



Figure 5: **Per-subject view of Table 4.** On steady-state responses four encoders fall below their patch embeddings (a); pretraining raises the encoder more than the patch embeddings in 14 of 20 pairs (b). ST-EEGFormer saw BCI IV-2a in pretraining.

Table 6: **Secondary endpoints at fixed rank**: pretrained patch embeddings / encoder (%).

| Model | Rank 16 | Rank 64 | Rank 256 |
|---|---|---|---|
| BCI IV-2a (BCA) | | | |
| EEGPT | 24.67 / 26.70 | 27.20 / 27.31 | 28.67 / 28.72 |
| CBraMod | 31.71 / 27.41 | 31.69 / 28.11 | 32.37 / 33.87 |
| LaBraM | 26.22 / 30.32 | 26.29 / 30.69 | 28.20 / 31.56 |
| LUNA | 30.30 / 28.55 | 32.75 / 31.75 | 34.99 / 32.85 |
| ST-EEGFormer | 24.98 / 33.08 | 26.14 / 32.43 | 26.79 / 36.86 |
| BNCI2015001 (BCA) | | | |
| EEGPT | 54.65 / 54.83 | 54.35 / 54.74 | 57.57 / 55.50 |
| CBraMod | 52.71 / 54.91 | 55.83 / 55.94 | 56.43 / 58.29 |
| LaBraM | 49.99 / 52.05 | 52.49 / 53.53 | 52.84 / 56.28 |
| LUNA | 53.60 / 53.34 | 53.70 / 53.84 | 58.29 / 55.18 |
| ST-EEGFormer | 55.57 / 58.85 | 54.86 / 62.11 | 55.15 / 64.89 |
| Nakanishi2015 (BCA) | | | |
| EEGPT | 59.57 / 62.90 | 68.46 / 70.74 | 80.49 / 71.42 |
| CBraMod | 11.11 / 23.33 | 65.43 / 35.12 | 71.17 / 46.67 |
| LaBraM | 21.54 / 19.81 | 64.01 / 41.79 | 74.07 / 53.21 |
| LUNA | 10.31 / 10.43 | 24.14 / 27.90 | 42.10 / 28.27 |
| ST-EEGFormer | 70.99 / 26.48 | 73.89 / 80.68 | 75.93 / 86.17 |
| ERP CORE P3 (AUROC) | | | |
| EEGPT | 59.84 / 58.93 | 64.81 / 62.54 | 70.84 / 65.98 |
| CBraMod | 65.25 / 59.46 | 72.56 / 65.26 | 76.06 / 72.25 |
| LaBraM | 63.96 / 62.56 | 68.24 / 69.95 | 73.49 / 72.50 |
| LUNA | 70.35 / 69.47 | 74.84 / 70.79 | 76.38 / 71.20 |
| ST-EEGFormer | 59.03 / 67.90 | 64.05 / 75.77 | 69.45 / 79.22 |

### B.2 A KNOWN LOCAL POWER CHANGE

This experiment compares learned representations with a fixed operator on a quantity whose true value is known (Section 3.3). We applied signed $\mu/\beta$ gains to left and right sensorimotor electrodes of 64 motor-imagery recordings from eight Lee2019 subjects not used for pretraining, together with two control changes, a common gain and a common phase (832 signals). Every representation shares 48 electrodes and a common average reference and is read through its own native sampling and scaling. Paired feature changes predict the injected log-power ratio under a subject-held-out ridge; feature scales are fitted on unperturbed training backgrounds (Table 7). Even an unfitted lateralisation contrast, $\ell_b = |L|^{-1} \sum_{c \in L} \log P_{c,b} - |R|^{-1} \sum_{c \in R} \log P_{c,b}$, reaches $R^2 = 0.9716$ with an error of about $2 \times 10^{-8}$ on the common-gain control, because a common gain adds the same constant to both terms. In paired subject bootstraps the complete encoders of CBraMod and EEGPT fall below the two-second operator by 0.060 [0.037, 0.091] and 0.092 [0.065, 0.130]. The experiment measures how readable one physiological feature is; motor-imagery accuracy is measured in Section 5.5.

Table 7: **A fixed power operator recovers a known power change best.** Control response: error on common-gain and common-phase changes, which inject no lateralised change.

| Frozen representation | $R^2 \uparrow$ | Control response $\downarrow$ |
|---|---|---|
| Raw waveform, linear readout | −0.0308 | 0.0438 |
| Fixed local power, 2 s | 0.9881 | 0.5682 |
| Fixed local power, 1 s | 0.9879 | 0.1913 |
| CBraMod patch embeddings | 0.6347 | 0.1099 |
| CBraMod random encoder | 0.4267 | 0.6414 |
| CBraMod pretrained encoder | 0.9282 | 0.2521 |
| EEGPT patch embeddings | 0.6126 | 0.7842 |
| EEGPT random encoder | 0.3043 | 1.9362 |
| EEGPT pretrained encoder | 0.8960 | 0.5998 |

### B.3 Ablation of the objective

This ablation asks whether changing only the objective changes what a model makes readable (Section 3.4). We continued one released CBraMod checkpoint, with both input branches intact. Twenty-four people with two recordings each are split into eight for the self-supervised continuation, four for a common structure readout and twelve for evaluation. Three paired continuation seeds share data, masks and a 64-step budget. The readout predicts the four-band tapered log-spectral shape from the same hidden states and is refitted at every evaluated checkpoint; intervals resample seeds and people independently (Table 8). In the pre-specified condition with the exact polarity expectation (Equation 9), the added structural target reduces the readout error by 0.0958 [0.0390, 0.1550] and improves over the initial checkpoint by 0.1096 [0.0835, 0.1356]; with the native target the reduction is 0.1118 [0.0621, 0.1612].

Table 8: **Adding a spectral-shape target makes the spectral shape more readable** (readout MSE, lower is better).

| Target condition | Waveform only | Waveform + spectral shape |
|---|---|---|
| Native | 0.7793 | 0.6675 |
| Exact polarity expectation, $\kappa = 0$ | 0.7297 | 0.6339 |
| Sampled polarity, $\kappa = 0$ | 0.7388 | 0.6330 |
| Initial checkpoint | 0.7435 | |

### B.4 Ablation of the input feature space

This ablation asks whether, at a fixed objective, changing only the input feature space changes what the same budget learns (Section 3.4). We held the architecture, initialisation, masks, data and 300 updates fixed; every token keeps its 200 waveform samples and receives four extra inputs, either fixed orthogonal random projections of the waveform or four tapered spectral-shape values, under the same waveform-plus-structure objective (Table 9).

Table 9: **Spectral-shape inputs make the spectral shape easier to learn at the same objective** (MSE, lower is better).

| Read out by | Random projections | Spectral shape | Reduction [95% interval] |
|---|---|---|---|
| Linear readout of the hidden states | 1.2395 | 0.8272 | 0.4123 [0.3316, 0.5111] |
| The model's own prediction | 1.0295 | 0.8121 | 0.2174 [0.1113, 0.3598] |

### B.5 Phase randomisation on frozen released models

This experiment tests whether the pretraining objective and classification rely on the same properties of the signal (Section 3.3). It reuses the BCI IV-2a and Nakanishi2015 cohorts of the patch-embedding analysis and their five outer folds. Each fold evaluates the pretrained patch embeddings, the pretrained complete encoder and a random complete encoder (one seed, distinct from the three seeds above). A ridge head is fitted on clean training trials with its previously selected penalty and held fixed for every condition; all model parameters stay frozen and subjects are weighted equally. The changes act after each model's native input preprocessing (200 Hz and microvolts divided by 100 for CBraMod; 256 Hz with per-trial exponential moving standardisation for EEGPT), so that the spectrum the model receives is the one preserved; the four-second inputs are then patched at one second and 0.25 s respectively. Phase draws depend on a fixed seed and the trial identity and never on labels. A common phase draws one value per non-DC, non-Nyquist frequency, shared across channels, on $\pm\pi/4$ (weak) or $\pm\pi$ (strong); independent phases keep each channel's power and change the cross-spectra; a removal condition zeroes 8–30 Hz; and a Fourier round trip with unchanged coefficients reproduces the clean predictions exactly.

The reconstruction error was measured on CBraMod's own reconstruction pathway: its released head and a 50% patch mask, identical per trial across conditions, with the loss on hidden patches only, each subject's error divided by the error of predicting zero and then averaged over subjects. CBraMod embeds each one-second patch in the time domain and through its Fourier magnitude,

and the four-second trials differ from its 30-second pretraining inputs. EEGPT's two pretraining objectives were not restored.

Under strong common phases ($\pm\pi$), the pretrained CBraMod encoder falls from 55.62% to 13.27% on Nakanishi2015 and from 42.44% to 31.23% on BCI IV-2a, and EEGPT from 71.98% to 16.67% and from 29.11% to 25.54%; CBraMod's relative reconstruction error moves from 0.4421 to 0.4141 and from 0.2644 to 0.2603, and rises to 1.2413 and 1.1745 under independent phases. In EEGPT's own ablation, removing the alignment objective changes the reconstruction loss from 0.56 to 0.57 while BCI IV-2a balanced accuracy falls from 58.46% to 52.87% (Wang et al., 2024). For the pretrained CBraMod encoder, weak common phases give 50.99% on Nakanishi2015 and 41.20% on BCI IV-2a, independent phases 12.16% and 26.23%, and removing 8–30 Hz 14.51% and 39.00%. Under strong common phases its patch embeddings fall from 77.53% to 21.60% and from 43.79% to 30.27%, and its random encoder from 47.47% to 12.72% and from 32.41% to 26.81%; EEGPT's random encoder falls from 71.48% to 15.00% and from 29.59% to 25.62%. EEGPT starts near chance on BCI IV-2a, which explains its smaller decline there. These are means without significance tests.

To confirm that the preserved power carries task information, a power reference reads the full-window $\log(1 + |\mathrm{rFFT}(X)|^2)$ of every channel from 1 to 40 Hz with a standardisation fitted on clean trials and a ridge head at a fixed relative penalty of one. Under CBraMod's preprocessing it scores 31.73% on BCI IV-2a and 82.53% on Nakanishi2015, under EEGPT's 26.83% and 52.04%; every phase condition and the round trip leave its trial-level predictions unchanged, and removing 8–30 Hz gives 25.58% and 37.96% (CBraMod preprocessing) and 25.93% and 16.30% (EEGPT). Its hyperparameters were selected separately, so it serves as a reference for the neural networks' declines.

Three facts bound what the experiment shows. The patch embeddings and the random encoders are about as sensitive as the pretrained encoders, so the dependence comes with the architecture. The phase change also alters the spectra of short patches and stimulus-locked timing, which can carry task information, notably for steady-state responses, and CBraMod reads frequency-domain as well as time-domain patches. The classifier was kept fixed, so its drop measures how a decision rule transfers. A direct test of what the encoders rely on needs a task in which the label is set by a known quantity and a temporal cue that accompanies it in training is decorrelated at test.

### B.6 Gradient direction at a released checkpoint

To see which structures the released objective pushes the encoder towards, we fitted, at the released CBraMod checkpoint, frozen readouts on eight BNCI2015001 subjects that read the marginal band power (P) and the cross-channel correlation structure beyond it (A | P) on two held-out subjects. The gradient of the released masked-reconstruction objective with respect to the 4,883,800 encoder parameters has cosine $+0.0279$ with the direction that improves the P readout and $-0.1289$ with the direction that improves the A | P readout, against $4.9 \times 10^{-5}$ for a matched random direction. A normalised step along the negative pretraining gradient improves the first readout and degrades the second at three displacements inside the first-order regime. The released objective thus favours marginal power over the cross-channel structure beyond it.

## C The five views in full

This appendix completes the description of Section 4.2 so that each view can be reproduced. All views share the same preprocessing: signals are calibrated, band-passed from 0.1 to 75 Hz with 50/60 Hz notch filters and resampled to 256 Hz before any operator; $x_t \in \mathbb{R}^C$ is one sample and $p_i \in \mathbb{R}^3$ the position of sensor $i$. All encoders have width 192 and eight attention heads. Masks act on the operator output before the encoder's input projection, so no gradient re-enters through hidden entries. Table 10 summarises the five views, Figure 6 illustrates them, Table 11 gives the full card and Figure 7 the five encoders.

**Waveform.** The target $Y_{c,n}$ of Equation 5 has 258 real values. The mask $\mathcal{M}$ hides 25% of the one-second patches of each sensor; on one-second inputs it hides two of the eight micro-patches

Table 10: **The five views of OpView**: what each operator exports and removes, and what its encoder predicts.

| View | Operator output | Removes | Native target | Hidden part |
|---|---|---|---|---|
| Waveform | Calibrated one-second patches | nothing | Whitened, tapered spectrum of hidden patches | 25% of patches per sensor |
| Frequency | Log power per sensor and 1 Hz group; phase consistency $r_b$ | per-sensor phase | Standardised log power of hidden sensors | about 25% of sensors per three groups, all windows |
| Relation | Gain-free cross-spectral shape $A_b$ and scale $e_b$ in a reference-free basis | reference; common gain (shape) | Rows of hidden anchor modes; the scale | 25% of anchor modes, all windows and bands |
| Projective | Polarity-free scalp state $Q_I$ and log amplitude per 62.5 ms token | sign; order within a token; detail beyond degree 3 | Hidden states and amplitudes | 25% of tokens, single or in spans of four |
| Dynamics | Level and eight causal differences | nothing | Change 62.5, 125 and 250 ms ahead | the future |




Figure 6: **The five views**: input, operator output, token and self-supervised target.

Table 11: **Architecture card.** $C$: sensors; $N_v$: time tokens of view $v$; $G = 75$ one-Hz groups. Parameter counts include the prediction and retention heads and the 16-token interface.

| View | Operator output | Encoder, target and retention | Params |
|---|---|---|---|
| Waveform | Calibrated one-second patches, common-average reference, each split into eight 32-sample micro-patches; grid $C \times N_W \times 256$. | Micro-patch stem $32 \to 64$ (depthwise convolution and within-patch attention), ordered concatenation to 192; four channel- and four time-attention blocks. Target: 129 complex bins of the whitened, Hann-tapered (floor 0.02) patch spectrum, per-bin power EMA (momentum 0.99); 25% mask, per sensor on multi-second inputs, and on one-second inputs two of eight micro-patches hidden per sensor or at the same times on every sensor. Retention: five-band power of visible patches. | 4.43 M |
| Frequency | Orthonormal real FFT, log power $\log(1 + \lvert F_{c,b}\rvert^2/s_b^2)$ with a per-bin scale fitted on training data; phase relative to the across-sensor mean phasor, set to zero for the encoder (the resultant $r_b$ enters only the operator features); 75 one-Hz groups over 1, 2 or 4 s supports; grid $C \times G \times N_F$. | Per-group projection to 192; sparse spectral graph of neighbouring groups, second and third harmonics in physical Hz and six band hubs; eight channel/frequency/time blocks and two coarse time blocks. Target: log power, standardised by training statistics, of about a quarter of the sensors hidden in each block of three groups over all time (visible cells weighted 0.5). Retention: five-band power. | 4.30 M |
| Relation | Contrast basis $W^\top \mathbf{1} = 0$; Welch band cross-spectral density $S_b$ of $W^\top X$ in six bands (0.5–4, 4–8, 8–13, 13–30, 30–45, 45–75 Hz); shape $A_b = \log(S_b/q_b + 10^{-3} I)$, scale $e_b = \log q_b$, $q_b = \operatorname{tr} S_b/K$; rank 3, 9 or 12 for 1, 2 or 4 s. | Node stem $1 \to 192$, complex-edge stem $2 \to 192$, geometry and band embeddings; two within-band graph attention stages, one across-band stage, two per-band temporal blocks. Target: Hermitian completion of $\operatorname{round}(K/4)$ hidden anchor modes (1, 2 or 3; at least two stay visible), visible block selected before normalisation; scale by a separate shallow head with no graph gradient. Retention: montage-mean power. | 4.97 M |
| Projective | Real spherical harmonics of degrees 1–3, $a_t = B^+ x_t \in \mathbb{R}^{15}$; per 16-sample token the energy-weighted projector mixture $Q_I$ (120 values) and log amplitude $e_I$. | Stem $121 \to 192$ with depthwise temporal convolution; 32 learned quadratic direction probes; eight temporal Transformer layers and a four-token coarse context. Target: masked state and amplitude (weights 0.9 and 0.1); 25% token mask, single tokens or spans of four. Retention: visible topography and past power. | 4.29 M |
| Dynamics | Level and eight causal differences $\Delta_k x_t = x_t - x_{t-k}$, $k \in \{1, \ldots, 128\}$ in powers of two; 16-sample per-sensor tokens of 144 values; grid $C \times N_D \times 9 \times 16$. | Per-field projection to 48 and lag attention with log-lag codes to 192; four fine and four coarse causal channel/time blocks; a coarse token becomes visible only after its four fine tokens. Target: $x_{t+h} - x_t$ for $h = 16, 32, 64$ samples, Smooth-L1 with fitted scales. Retention: band power of the previous two seconds. | 4.36 M |

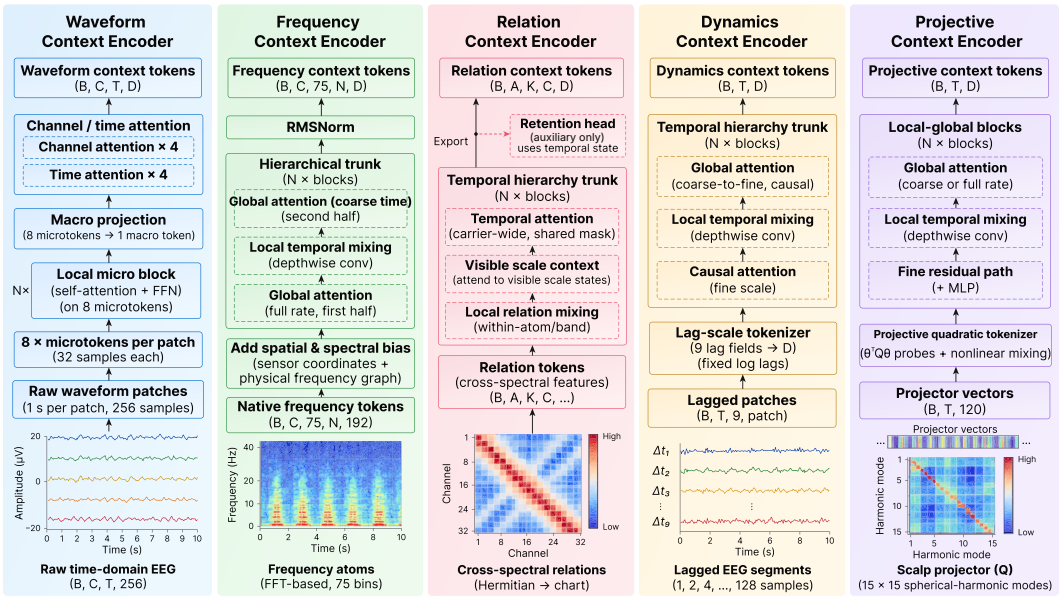



Figure 7: **The five encoders** (schematic). $B$: batch; $C$: sensors; $T$: time tokens; $D$: width.

instead, either per sensor or at the same times on every sensor, with equal probability, and

$$L_W = \frac{1}{|\mathcal{M}|} \sum_{(c,n)\in\mathcal{M}} \|\hat{Y}_{c,n} - Y_{c,n}\|^2 + \frac{1}{2}\frac{1}{|\mathcal{M}^c|} \sum_{(c,n)\notin\mathcal{M}} \|\hat{Y}_{c,n} - Y_{c,n}\|^2 + \lambda R_W. \tag{11}$$

**Frequency.** Equation 6 uses the non-DC bins through 75 Hz, grouped into 75 one-Hz groups. The resultant $r_b$ replaces the $C$ per-sensor phase angles by one scalar per bin and removes the common same-frequency rotation, the declared invariance. The mask $\mathcal{M}$ splits the 75 groups into 25 blocks of three and, in each block, hides about a quarter of the sensors for every time window of the recording, leaving at least two and a majority of sensors visible. With $\tilde{P}_{c,b,n} = (P_{c,b,n} - \mu_b)/\sigma_b$ standardised by per-bin statistics fitted on training data and $\mathcal{V}$ the visible cells,

$$L_F = \frac{1}{|\mathcal{M}|} \sum_{\mathcal{M}} \big(\hat{P} - \tilde{P}\big)^2 + \frac{1}{2}\frac{1}{|\mathcal{V}|} \sum_{\mathcal{V}} \big(\hat{P} - \tilde{P}\big)^2 + \lambda R_F. \tag{12}$$

**Relation.** Welch averaging with a periodic Hann window and 50% overlap, over segments of 0.5, 1 and 4 s for supports of 1, 2 and 4 s (a single tapered segment at 4 s), gives $S_b = \mathbb{E}[F_b F_b^*] \in \mathbb{C}^{K\times K}$ in each of six bands. Nodes are the diagonal of $A_b$ and edges its off-diagonal entries, real and imaginary. The columns of $W$ are smooth spatial functions around fixed anchor positions, and the masked units are these anchor modes: the objective hides a set $H_b$ of modes and completes their rows,

$$L_R = \frac{1}{6}\sum_b \Big( \|\hat{A}_b[H_b] - A_b[H_b]\|_F^2 + (\hat{e}_b - e_b)^2 \Big) + \lambda R_R. \tag{13}$$

To keep hidden entries out of the input, the visible block is selected before the trace normalisation and the matrix logarithm; the decoder is Hermitian, and the scale head sends no gradient into the graph encoder. The objective hides $\mathrm{round}(K/4)$ anchors, keeping at least two visible, and every time window and band of a recording shares the same hidden anchors, so a neighbouring window cannot expose their measurements. For recordings in bipolar montages, such as CHB-MIT, the same spatial functions are applied to the bipolar derivations, and the view reads the part of the signal that they observe.

**Projective.** With $r_t = \|a_t\|/\sqrt{4\pi}$ and $v_t \in \{0, 1\}$ requiring an average-referenced global field power of at least 2 $\mu$V and at least 0.8 explained energy, the token of Equation 8 is, in full,

$$Q_I = \frac{\sum_{t\in I} v_t r_t^2\, \vartheta_t \vartheta_t^\top}{\sum_{t\in I} v_t r_t^2}, \qquad e_I = \tfrac{1}{2}\log \frac{\sum_{t\in I} v_t r_t^2}{\sum_{t\in I} v_t}. \tag{14}$$

$Q_I$ has unit trace and a rank that can exceed one; it is exported as 120 symmetric values ($\sqrt{2}$ on off-diagonals, preserving the Frobenius norm) together with $e_I$. The objective weights states and amplitudes 0.9 and 0.1 under a 25% token mask, half the updates hiding single tokens and half spans of four, with at least one visible and one hidden token per row.

**Dynamics.** The lags are $1, 2, 4, \ldots, 128$ samples, and the fields are grouped into 16-sample per-sensor causal tokens. The objective is

$$L_D = \sum_{h\in\{16,32,64\}} \frac{1}{s_h}\, \mathrm{SmoothL1}\big(\hat{\Delta}_t^{(h)},\, x_{t+h} - x_t\big) + \lambda R_D, \tag{15}$$

with scales $s_h$ fitted on training data, and undefined prefixes and terminal targets excluded. A coarse token pools four fine tokens and becomes visible only after all four have been observed, and the retention head reads only the previous two seconds. The operator and the encoder are causal with respect to the pretraining recordings, which are band-passed and notch-filtered with zero-phase filters, so each sample already carries the filter's short two-sided support.

**Retention and the exported interface.** Two auxiliary terms accompany every native objective. The first is a band-power retention readout, scaled so that it cannot dominate the native loss, $L_v = L_{v,\mathrm{native}} + 0.2\,\mathrm{sg}(T_{v,\mathrm{native}})\, R_{v,\mathrm{power}}/\mathrm{sg}(D_v)$, where $T_{v,\mathrm{native}}$ is the native reference loss and sg stops gradients. The second trains a compact interface: each view exports its encoder's output tokens,

and sixteen learned queries also cross-attend to those tokens to give $Z_v \in \mathbb{R}^{16\times 192}$, trained with weight 0.05 to reconstruct the detached, normalised tokens from their physical positions alone. The interface adds about 60k parameters per view and leaves the full tokens available. Relation detaches its shape and scale states before the interface, so the two keep separate gradients. Neither term was ablated separately from the native objective, so the gains we report are those of the complete objective. Views are trained independently so that each can be added or replaced alone; we have not compared this with joint training.

# D PRETRAINING DATA, OPTIMISATION AND THE PRETRAINING TASKS

All encoders in this paper are pretrained on the same public corpus and differ only in the number of updates, 40,000 or 160,000. This appendix describes the corpus, what longer training changes, the optimisation, and the evaluation of the pretraining tasks in Section 5.2.

**Corpus.** Each view is pretrained once, with one seed, on the sources of Table 12 whose channel count and duration its operator can process, 52 to 63 sources depending on the view, with durations computed after subject splits. Sources are drawn by fixed weights that raise small datasets, and the 40,000-update and 160,000-update runs share the objective, optimiser and batch rule; the 160,000-update Relation run also masks by region. Besides the sources in the table, the corpus holds the unlabelled recordings of the pretraining-split subjects of seven evaluation datasets, BCI IV-2b, BNCI2015001, BNCI2014008, BNCI2014009, CHB-MIT, SEED and Nakanishi2015 (715 hours for Waveform and Dynamics, 50–52 hours for the other views), which are therefore marked ‡ in the tables; BCI IV-2a and EEGMat are absent from the corpus.

**Longer training.** To see what longer training changes, we compared the 160,000-update and 40,000-update encoders on subjects who took no gradient step in pretraining. Frozen across subjects under the benchmark's probe, the change is $+1.4$ points on average for Waveform (better on 6 of 9 datasets) and $+1.6$ for Dynamics (8 of 9), and $+0.3$, $-0.3$ and $-0.9$ for Frequency, Projective and Relation; the longer Relation run also masks by region, so its change is not due to length alone. Longer training thus strengthens what an objective already rewards (Proposition 1). On BCI IV-2a, absent from the corpus and from model development, the Dynamics encoder rises from 36.03% to 40.90% frozen, against 37.01% for its matched random encoder; these are means over all nine subjects and three classifier seeds, with unchanged output dimensions and no test-subject alignment.

**Encoders behind each result.** The 160,000-update Waveform and Dynamics encoders produce the frozen single-view results across subjects without alignment (Table 18 and the corresponding View entries of Table 2), the within-subject probe results on BNCI2014008 and BNCI2014009, with and without alignment, and the fine-tuned Dynamics result and the frozen multi-view result on BCI IV-2a. All other results use the 40,000-update Waveform and Dynamics encoders: fine-tuning across subjects and within-subject results on the other datasets, the other aligned and multi-view results, the controlled comparisons of Tables 1 and 17, the pretraining tasks of Section 5.2 and the diagnostics of Appendix H. The 160,000-update runs of these settings could not be completed in time and will be added. The Frequency, Relation and Projective views use their 40,000-update encoders throughout; longer training changed their frozen results by less than one point.

**Optimisation.** All views use AdamW with weight decay 0.05, gradient clipping at 1, bfloat16 computation and a linear warm-up followed by cosine decay; the cosine schedules of the Frequency and Dynamics views span 80,000 updates, and the Waveform view uses at most 16 records per batch. Parameter counts are in Table 11.

## D.1 THE PRETRAINING TASKS ON SUBJECTS WITHOUT GRADIENT STEPS

Section 5.2 checks whether each encoder learned the structure its objective rewards. To answer this on data the encoders had not trained on, we ran the 40,000-update encoders, without any update, on development subjects of three corpus datasets who took no gradient step: six ERP CORE flanker subjects with one-second trials for the Waveform and Projective views, eight Lee2019 motor-imagery subjects with four-second trials for the Waveform, Relation and Dynamics views, and eight Lee2019

Table 12: **Sources of the pretraining corpus** (hours per view; W/D: Waveform and Dynamics; F: Frequency; R: Relation; P: Projective; –: the operator does not process the source). Evaluation datasets are described in the text.

| Source | W/D | F | R | P |
|---|---|---|---|---|
| *Motor imagery / execution* | | | | |
| Cho2017 | 18.1 | 18.1 | 18.1 | 18.1 |
| DICIS MI | 1.0 | – | – | 1.0 |
| Dong2023 | 9.0 | – | – | – |
| ds003810 | 3.1 | – | 3.1 | 3.1 |
| ds005342 | 22.7 | 22.7 | 22.7 | 22.7 |
| EEGMMIDB | 28.2 | 28.2 | 28.2 | 28.2 |
| Guttmann-Flury2025 ME | 5.0 | 5.0 | 5.0 | 5.0 |
| Guttmann-Flury2025 MI | 9.9 | 9.9 | 9.9 | 9.9 |
| Lee2019 MI | 64.4 | 64.4 | 64.4 | 64.4 |
| MI-BMPI | 1.3 | – | – | 1.3 |
| MILimb | 5.6 | – | 5.6 | 5.6 |
| Schirrmeister2017 | 19.9 | 19.9 | 19.9 | 19.9 |
| Stieger2021 | 399.2 | 399.2 | 399.2 | 399.2 |
| *ERP / P300 / ErrP* | | | | |
| BNCI2015-013 | 4.0 | 4.0 | 4.0 | 4.0 |
| ds003825 | 29.5 | 29.5 | 29.5 | 29.5 |
| ds003846 | 15.3 | 15.3 | 15.3 | 15.3 |
| ds004262 | 5.8 | 5.8 | 5.8 | 5.8 |
| EPFL P300 | 2.2 | 2.2 | 2.2 | 2.2 |
| ERP CORE flanker | 6.2 | 6.2 | 6.2 | 6.2 |
| ERP CORE MMN | 4.7 | 4.7 | 4.7 | 4.7 |
| ERP CORE N170 | 4.5 | 4.5 | 4.5 | 4.5 |
| ERP CORE N2pc | 4.8 | 4.8 | 4.8 | 4.8 |
| ERP CORE N400 | 3.5 | 3.5 | 3.5 | 3.5 |
| ERP CORE P3 | 3.1 | 3.1 | 3.1 | 3.1 |
| Guttmann-Flury2025 P300 | 7.9 | 7.9 | 7.9 | 7.9 |
| Lee2019 ERP | 40.9 | 40.9 | 40.9 | 40.9 |
| ReGAIT start | 7.0 | 7.0 | 7.0 | 7.0 |
| ReGAIT stop | 6.4 | 6.4 | 6.4 | 6.4 |
| *SSVEP / c-VEP* | | | | |
| BETA (Liu2020) | 8.4 | – | 8.4 | 8.4 |
| Guttmann-Flury2025 SSVEP | 2.8 | 2.8 | 2.8 | 2.8 |
| Lee2019 SSVEP | 57.2 | 57.2 | 57.2 | 57.2 |
| MAMEM1 | 3.8 | 3.8 | 3.8 | 3.8 |
| MAMEM2 | 3.5 | 3.5 | 3.5 | 3.5 |
| MAMEM3 | 3.2 | – | 3.2 | 3.2 |
| Thielen2021 | 20.1 | 20.1 | 20.1 | 20.1 |
| Wang2016 | 10.0 | – | 10.0 | 10.0 |

| Source | W/D | F | R | P |
|---|---|---|---|---|
| *Emotion* | | | | |
| ds002721 | 17.9 | 17.9 | 17.9 | 17.9 |
| ds003004 | 34.8 | 34.8 | 34.8 | 34.8 |
| ds004315 | 14.2 | 14.2 | 14.2 | 14.2 |
| FACED | 109.4 | 109.4 | 109.4 | 109.4 |
| SEED-IV | 30.6 | 30.6 | 30.6 | 30.6 |
| SEED-V | 31.9 | 31.9 | 31.9 | 31.9 |
| *Resting / mental state* | | | | |
| Cattan2019 PHMD | 1.8 | 1.8 | 1.8 | 1.8 |
| ds001787 | 18.9 | 18.9 | 18.9 | 18.9 |
| ds002778 | 1.7 | 1.7 | 1.7 | 1.7 |
| ds003775 | 7.1 | 7.1 | 7.1 | 7.1 |
| ds003969 | 46.0 | 46.0 | 46.0 | 46.0 |
| ds004148 | 52.5 | 52.5 | 52.5 | 52.5 |
| ds004572 | 36.6 | 36.6 | 36.6 | 36.6 |
| ds004902 | 12.1 | 12.1 | 12.1 | 12.1 |
| Hinss2021 resting | 0.6 | 0.6 | 0.6 | 0.6 |
| *Cognitive task* | | | | |
| ds002680 | 14.7 | 14.7 | 14.7 | 14.7 |
| ds003690 | 32.9 | 32.9 | 32.9 | 32.9 |
| ds004279 | 38.4 | 38.4 | 38.4 | 38.4 |
| ds005095 | 12.0 | 12.0 | 12.0 | 12.0 |
| Hinss2021 workload | 8.2 | – | – | 8.2 |
| *Sleep* | | | | |
| ds005620 | 5.9 | 5.9 | 5.9 | 5.9 |
| HMC sleep | 792.7 | 792.7 | – | – |
| Sleep-EDFx cassette | 2380.7 | – | – | – |
| Sleep-EDFx telemetry | 266.3 | – | – | – |
| *Clinical (epilepsy)* | | | | |
| Siena Scalp EEG | 93.8 | 93.8 | 93.8 | 93.8 |
| *Mobile / motion* | | | | |
| Lee2021 mobile | 21.5 | 21.5 | 21.5 | 21.5 |
| PhysioMotion (ds006386) | 42.5 | 42.5 | 42.5 | 42.5 |
| Total | 4,968 | 2,271 | 1,509 | 1,519 |
| Sources | 63 | 52 | 56 | 59 |

steady-state subjects with four-second trials for the Frequency view. To rule out a favourable choice of trials, we fixed the trials before inference, the first two of every subject and class (three for Relation), and drew masks by each view's own pretraining rule.

Each case is scored by the encoder's pretraining error on the masked part, divided by the error of the best simple baseline that sees only the visible part: interpolation in time or from the three nearest visible electrodes for the Waveform view, the three nearest visible electrodes at the same frequency for the Frequency view, the training mean of the target for the Relation view, the visible mean or time interpolation for the Projective view, and persistence, a forecast of no change, for the Dynamics view. A matched random encoder, with the same architecture and training statistics and re-initialised weights, is 1.2 to 43 times worse than these baselines, so the encoders' advantage comes from pretraining. The summary pools 64 Frequency, 80 Waveform, 48 Relation, 48 Projective and 32 Dynamics cases, the Dynamics trials holding 109,120 forecasts. Figure 8 shows the case with the lowest ratio of each view, Figure 3a summarises all cases, and Table 13 lists every task.

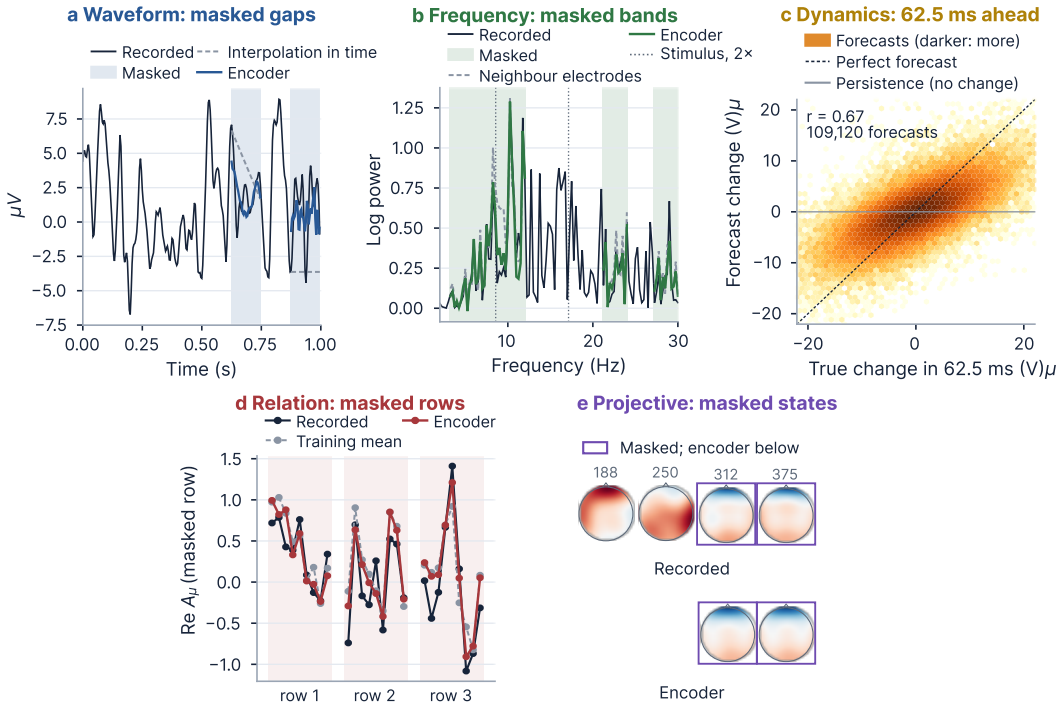



Figure 8: **Best case of each pretraining task**: the encoder follows the recording more closely than the best simple baseline (the case with the lowest error ratio of each view).

Table 13: **Every encoder beats its best visible-only baseline on its main task** (median error ratio; below one, the encoder is better).

| View | Task | Best visible-only baseline | Error ratio |
|---|---|---|---|
| Dynamics | Forecast the change 62.5 ms ahead | Persistence (no change) | 0.55 |
| | 125 ms ahead | | 0.59 |
| | 250 ms ahead | | 0.61 |
| Frequency | Complete masked electrode-bands of log power | Nearest visible electrodes | 0.77 |
| Waveform | Fill masked 125 ms gaps | Interpolation in time or space | 0.85 |
| | gaps shared by every electrode | | 1.03 |
| | masked one-second patches | | 0.97 |
| Relation | Complete masked rows of the cross-spectral shape | Training mean | 0.87 |
| Projective | Complete masked scalp states | Visible mean or interpolation | 0.93 |
| | masked 250 ms spans | | 1.05 |

## E EVALUATION PROTOCOL

This appendix shows that every single-view result follows the protocol of EEG-FM-Compass (Liu et al., 2026b). It states the protocol and ends with a list of every departure from it.

**Datasets and settings.** The nine datasets span motor imagery (BCI IV-2a, IV-2b (Tangermann et al., 2012), BNCI2015001), P300 (BNCI2014008, BNCI2014009), steady-state responses (Nakanishi2015 (Nakanishi et al., 2015)), mental arithmetic (EEGMat (Zyma et al., 2019)), emotion (SEED (Zheng & Lu, 2015)) and seizure detection (CHB-MIT (Shoeb, 2009; Goldberger et al., 2000)). We use the benchmark's subject and usage splits and its four settings: frozen probing and full fine-tuning across subjects, with one subject left out (LOSO), and both within subject on its calibration split. The pretraining corpus contains no recording of BCI IV-2a or EEGMat. Euclidean alignment is applied only where marked and compared with EEGPT (Wang et al., 2024) and MIRepNet (Liu et al., 2026a).

**Benchmark protocol.** The benchmark's linear probe is dropout and a linear layer trained with AdamW at $10^{-3}$ for 20 epochs and scored at the final epoch; full fine-tuning uses an encoder learning

rate of $10^{-4}$. We apply both to a view's exported tokens. Where the exported features are very wide, a fixed, label-free count-sketch projection reduces them before the probe: to 1,024 columns on BNCI2014008 and 4,096 on EEGMat, SEED and CHB-MIT across subjects, and to 4,096 for the Frequency view on SEED within subject.

**Windows and splits.** MOABB windows are offsets from each dataset's registered interval. BCI IV-2b is read at 3 to 7 s; BNCI2015001, BNCI2014008 and BNCI2014009 are read unshifted; Nakanishi2015 windows are shifted by 0.15 s and stay inside the stimulation period. For datasets with a usage column (CHB-MIT, SEED), a leave-one-subject-out fold trains on the other subjects' `train` rows and tests on the held-out subject's `test` rows.

**Alignment.** Results are split by alignment. Without alignment, single feature sets reach the benchmark's probe as exported, as for every published probe, and the closed-form ridge readouts of Appendix F standardise each column on the training rows. With Euclidean alignment (He & Wu, 2020), frozen features are standardised per subject and fine-tuning inputs are whitened per subject, each using that subject's own unlabelled trials, the held-out subject included. Aligned results are therefore transductive; they are compared only with the benchmark's two models that use Euclidean alignment. Within-subject fine-tuning uses a scale-preserving form, $x \mapsto (\operatorname{tr} \bar{R}/C)^{1/2}\, \bar{R}^{-1/2} x$ for a subject's mean covariance $\bar{R}$ over $C$ channels, because whitening to unit covariance shrinks the input to about 1 $\mu$V, far below the amplitudes of pretraining.

**Published rows.** Published rows are parsed from the EEG-FM-Compass tables by code. For each dataset and setting we compare with the strongest of the twelve foundation models that the benchmark evaluates. The benchmark applies Euclidean alignment only to EEGPT and MIRepNet (Liu et al., 2026a), the models whose original pipelines use it; EEGPT is among the twelve, and MIRepNet, which appears in the benchmark's per-dataset motor-imagery tables (accuracy on balanced classes), enters the aligned comparison. SEED's published rows are plain accuracy; on its near-balanced classes our accuracy and balanced accuracy differ by less than 0.2 points where both were recorded. The Trad. entries of Table 2 are the benchmark's traditional decoders, trained from scratch under its protocol in the two fine-tuning settings, the only ones in which it reports them: CSP with linear discriminant analysis on the motor-imagery datasets, EEGMat and SEED, xDAWN with linear discriminant analysis on BNCI2014008 and BNCI2014009, TRCA on Nakanishi2015 and power-spectral features with a support vector machine on CHB-MIT.

**Comparability with EEG-FM-Compass.** *Identical to the benchmark:* the nine datasets and their labels, the subject and usage splits, the four settings, the probe and fine-tuning protocol above for every single feature set and fine-tuned view, the metric, the comparison set, the separate treatment of alignment, and the ‡ mark for a model that saw an evaluation dataset in pretraining, which we apply to our own rows (Appendix D). Every departure is listed in Table 14 with its mark in Table 2.

**The Fusion column.** The View column reads each pretrained view exactly as the benchmark reads a published model; the Fusion column departs only in how several views are read, and five facts keep it comparable. First, it uses the benchmark's datasets, subject and usage splits, settings and metric, and no test label: every weight, penalty and choice is fitted on training subjects or calibration trials. Second, its readouts are linear classifiers of the views, the family of the benchmark's probe, or combinations of such classifiers' outputs; none learns new features from the signal. Third, the benchmark defines no readout for several representations, and its probe, applied to their concatenation, is limited by its fixed training budget: on BCI IV-2a it scores 33.3% on the five operator feature sets together, near the 25% chance level (Appendix F). Fourth, the closed-form ridge does not favour pretrained features: on EEGMat, where every view was read both ways, the pretrained encoders score within $-2.5$ and $+1.4$ points of their result under the benchmark's probe, while the random encoders gain 1.6 to 6.2 points (Appendix F.4). Fifth, the FM column takes, in each cell, the best of twelve published models, a choice made with the test results, whereas no Fusion entry uses a test label. The procedure is not identical across datasets (Appendix F). Frozen without alignment, BNCI2014008 and SEED use the weighted ridge alone, and CHB-MIT chooses only between the joint readout and single feature sets; frozen with alignment, two feature sets built from the held-out subject's unlabelled trials enter on BCI IV-2a and Nakanishi2015. After fine-tuning across subjects, the candidates also include four aligned operator feature sets on BCI IV-2a with alignment, and the Waveform operator

read by a trained head on BNCI2014008 without alignment. Within subject, the view for the probe entries of BNCI2014008 and BNCI2014009 is chosen by the benchmark's probe in place of the ridge. Because the Fusion readouts also draw on the operator and random-encoder features of each view, which need neither pretraining nor labels, they measure the whole system; the View column isolates the pretrained encoders.

Table 14: **Departures from EEG-FM-Compass.**

| Departure | Mark in Table 2 | Details |
|---|---|---|
| Readouts over several views are fitted separately, by closed-form ridge or by choosing a view on calibration trials; published models were read with the benchmark's probe alone, so the View column is the like-for-like comparison | Fusion column | Appendix F |
| The result comes from one classifier seed instead of three; the other two seeds could not be completed in time and will be added | $^{s}$ | Appendix F.1 |
| Without alignment, the rule's choice for the fine-tuned BNCI2014008 Fusion entry is the Waveform operator read by a trained head | $^{o}$ | Appendix F.2 |
| CHB-MIT inputs hold 27,436 windows against the benchmark's 29,840 | $^{*}$ | – |
| Cells cover the first seven held-out subjects; the remaining subjects could not be run in time and will be added | italic | – |
| SEED's published rows are plain accuracy | – | Published rows |
| Wide encoder features pass a fixed count-sketch projection | – | Benchmark protocol |

# F READOUTS THAT USE SEVERAL VIEWS

This appendix explains how the views are used together in Section 5.5 and why the results remain comparable with the benchmark. EEG-FM-Compass reads a foundation model through one probe trained on its embedding, and each OpView view, a complete pretrained encoder, is read exactly this way. The benchmark defines no readout for several representations at once. A joint head that contains every single-view head can do no worse, in population, than the best single view (Appendix A.5), provided it can be fitted; under its fixed training budget, the benchmark's probe reads concatenated feature sets poorly, scoring 33.3% on the 17,690 columns of the five operator feature sets of BCI IV-2a, near the 25% chance level. We therefore fit readouts for several views separately. They see only training subjects or calibration trials and change only how views are combined; everything else is the benchmark's (Table 14).

## F.1 FROZEN, ACROSS SUBJECTS

The frozen rule combines, for every view whose operator is defined on the dataset, the operator, random-encoder and pretrained-encoder feature sets, with fixed per-trial variants of some operators; all are label-free. Two feature sets are built from the held-out subject's unlabelled trials: the Relation operator of BCI IV-2a recentred on each subject, and Relation encoder features of Nakanishi2015 normalised across subjects. Both are excluded without alignment, so every result without alignment uses training-side statistics only; with alignment, itself transductive, both are admitted. For lack of time, the BCI IV-2a result without alignment has one classifier seed so far ($^{s}$ in Table 2); the other two will be added. In seven of its nine folds the inner split chose the weighted joint readout. The readout is a class-balanced ridge onto one-hot targets, with class weights inverse to class frequency. Each feature set is standardised by column on the training rows, scaled by the inverse square root of its width and weighted by $w_v = \max(0, \mathrm{BAC}_{\mathrm{inner}}(v) - 1/K)$, normalised to sum to one, where $\mathrm{BAC}_{\mathrm{inner}}(v)$ is the balanced accuracy of feature set $v$ alone under a leave-one-training-subject-out split. The penalty $\lambda = a \cdot \mathrm{mean\,diag}(K_{\mathrm{train}})$, $a \in \{10^{-3}, 10^{-2}, 10^{-1}, 1, 10\}$, is chosen on the same split, taking the largest penalty within 0.15 points of the best. In each fold the weighted joint readout competes with a selection under the benchmark's probe, whose candidates are every single feature set and stacked, convex and top-$k$ combinations of feature-set log-probabilities, a combination replacing the best single set only when its paired inner advantage exceeds one standard error; the fold keeps whichever the inner split scores higher, and the selection when the two lie within 0.15 points. On BNCI2014008 and SEED without alignment the result is the weighted ridge alone, and on CHB-MIT without alignment each fold chooses, on training subjects, between the joint readout and single

feature sets. The joint readout's weights and penalty are chosen on the inner split that scores it, which slightly favours it in this choice. Table 15 sets the result beside the best single feature set held fixed across folds, a reference chosen with hindsight.

Table 15: **Frozen multi-view rule and the best single feature set chosen with hindsight** (balanced accuracy, %; marks as in Table 14).

| | Without alignment | | | With alignment | | |
|---|---|---|---|---|---|---|
| Dataset | Rule | Best set | | Rule | Best set | |
| BCI IV-2a | 37.77$^{s}$ | – | | 53.33 | 50.00 | Rel. operator, recentred |
| BCI IV-2b‡ | 77.48 | 78.61 | Dyn. pretrained | 79.08 | 79.38 | Dyn. pretrained |
| BNCI2015001‡ | 63.38 | 63.50 | Dyn. pretrained | 71.33 | 68.25 | Dyn. pretrained |
| BNCI2014008‡ | 67.34 | 69.39 | Wave. operator | 68.31 | – | |
| BNCI2014009‡ | 69.65 | 74.33 | Wave. random | 76.89 | 76.16 | Wave. operator |
| Nakanishi2015‡ | 87.65 | 87.53 | Dyn. random | 88.64 | 88.64 | Dyn. random |
| EEGMat | 75.65 | 73.15 | Freq. pretrained | 86.39 | 85.37 | Freq. pretrained |
| SEED‡ | 47.70 | 49.75 | Freq. pretrained | 60.26 | 58.98 | Freq. pretrained |
| CHB-MIT$^{*}$‡ | 70.23 | 67.70 | Wave. pretrained | 77.12 | 73.79 | Wave. operator |

### F.2 Fine-tuned, across subjects

After fine-tuning, the views are combined by their class logits. Each view's encoder is fine-tuned separately under the benchmark's protocol, and the combination is a closed-form ridge over the views' class logits whose weights are fitted on training subjects only. To obtain logits for training subjects that the fine-tuned encoders have not seen, each view is also fine-tuned on one half of the training subjects to score the other half. The candidate weightings are nonzero on all views, on the top-$k$ views or on a single view, and a stacker trained with the probe is a further candidate; the inner split chooses among them in every fold. A weighting that is nonzero on one view alone is that view, so the rule selects a single view whenever the inner split finds that combining does not help. With alignment on BCI IV-2a, the candidates also include four aligned operator feature sets, and the combination reaches 60.0 against 59.5 for the fine-tuned Dynamics encoder alone. On BNCI2014008 without alignment, the candidates also include the Waveform operator read by a trained head, which the rule selects ($^{o}$ in Table 2). The aligned BNCI2014008 fine-tuning entries in both the View and Fusion columns report the same single Waveform view (40,000 pretraining updates, all eight subjects, one classifier seed), not a separate multi-view combination.

### F.3 Within subject

Within subject, each subject has a few dozen calibration trials, so the rule keeps one view per subject. For each view with a pretrained within-subject result, a class-balanced closed-form ridge is fitted on that view's frozen pretrained features using only the subject's calibration trials and scored by stratified three-fold cross-validation inside them, with the penalty grid and tie rule above. The view with the highest inner score is selected, ties within 0.15 points broken in the fixed order Dynamics, Frequency, Waveform, Relation, Projective, and the subject's score is that view's fine-tuned or probe score. No test trial enters any step (Table 16). Averaged over the nine datasets, the rule's fine-tuned score is 69.27, against 66.64 for always using the Dynamics view, the only view defined on every dataset, and 71.41 for the best view of each dataset chosen with hindsight, all from the seed-0 runs that the rule reads. The rule thus recovers about half of what hindsight would add; its largest miss is EEGMat, 9.5 points below the Relation view. Under the probe the rule gives 67.23, the Dynamics view 66.47 and the best view of each dataset 69.05, with single views from Table 20. On BNCI2014008 and BNCI2014009 the probe entries use the 160,000-update Waveform and Dynamics encoders, and the view is chosen by the benchmark's probe on the same calibration folds in place of the ridge; no test trial enters the choice.

### F.4 One readout rule on every view

Section 5.3 reads every view of six datasets with its operator, random-encoder and pretrained-encoder features under one readout rule fixed before the runs. Each feature set is read alone by the class-balanced closed-form ridge of Appendix F.1, with the same column standardisation, penalty grid and

Table 16: **Choosing a view on calibration trials recovers about half of the gap to the best view after fine-tuning** (balanced accuracy, %; single views fine-tuned, seed 0; best view chosen with hindsight).

| Dataset | Rule, fine-tuned | Rule, probe | Best view | Freq. | Wave. | Rel. | Dyn. | Proj. |
|---|---|---|---|---|---|---|---|---|
| BCI IV-2a | 42.32 | 37.20 | 48.80 | 34.48 | 31.10 | 48.80 | 34.64 | 27.67 |
| BCI IV-2b‡ | 77.91 | 76.79 | 76.26 | 74.24 | 63.79 | n/a | 76.26 | n/a |
| BNCI2015001‡ | 63.81 | 60.73 | 65.54 | n/a | 56.79 | n/a | 60.42 | 65.54 |
| BNCI2014008‡ | 54.22 | 58.23 | 55.43 | 50.28 | 53.12 | 51.91 | 55.43 | 52.39 |
| BNCI2014009‡ | 59.16 | 62.09 | 59.02 | 50.86 | 58.20 | 51.83 | 59.02 | 52.22 |
| Nakanishi2015‡ | 94.75 | 93.93 | 94.75 | 84.57 | 94.75 | – | 90.43 | 12.65 |
| EEGMat | 86.11 | 85.42 | 95.60 | 87.73 | 70.14 | 95.60 | 77.31 | 70.37 |
| SEED‡ | 62.41 | 55.90 | 62.67 | 62.41 | 59.75 | 44.66 | 62.67 | n/a |
| CHB-MIT‡ | 82.70 | 74.80 | 84.61 | 84.61 | 72.81 | n/a | 83.60 | n/a |
| Mean | 69.27 | 67.23 | 71.41 | | | | 66.64 | |

tie rule, the penalty chosen by a leave-one-training-subject-out split; the encoders are those of 40,000 updates at native width. The held-out subject enters no step without alignment; with alignment, each subject's features are standardised with its own unlabelled trials. Views not defined on a dataset are omitted, and BNCI2014008, SEED and CHB-MIT were not run.

Pretraining improves on the random encoder in 15 of 26 view–dataset pairs without alignment and in 18 of 26 with alignment (Table 17). The Frequency encoder improves in all ten pairs, by up to 10.5 points, the Projective encoder in eight of ten and the Dynamics encoder in eight of twelve, by up to 11.9 points; the Waveform encoder improves in four of twelve and the Relation encoder in three of eight. On EEGMat, the ridge reads the random encoders 1.6 to 6.2 points higher than the benchmark's probe (Table 22), while the pretrained encoders move by $-2.5$ to $+1.4$, so their margin over the random encoders shrinks from 4.6–9.0 points to 0.6–3.9 without alignment. The same rule also combines the three views that score best on the training subjects; combined, the random views already hold most of what this readout uses, with means over the six datasets of 65.8, 68.5 and 68.9 for the operator, random and pretrained features without alignment and 71.6, 72.9 and 72.9 with alignment, so our prediction that the pretrained features would lead the combination with alignment was not met.

Table 17: **Read one view at a time under one readout rule, pretraining improves on the random encoder for most views, the Frequency encoder in every pair** (frozen, across subjects; balanced accuracy, %; Op.: operator; Rand.: random encoder; Pre.: pretrained encoder; gain: Pre. minus Rand.).

| | | Without alignment | | | | With alignment (transductive) | | | |
|---|---|---|---|---|---|---|---|---|---|
| Dataset | View | Op. | Rand. | Pre. | Gain | Op. | Rand. | Pre. | Gain |
| BCI IV-2a | Freq. | 27.9 | 26.2 | 28.0 | +1.8 | 32.3 | 30.1 | 36.2 | +6.1 |
| | Wave. | 31.2 | 38.7 | 33.1 | −5.6 | 33.0 | 39.8 | 34.4 | −5.4 |
| | Rel. | 38.0 | 38.0 | 36.9 | −1.1 | 48.4 | 46.1 | 42.9 | −3.2 |
| | Dyn. | 36.3 | 37.2 | 38.4 | +1.2 | 41.4 | 38.4 | 42.4 | +3.9 |
| | Proj. | 31.5 | 28.7 | 31.7 | +3.0 | 36.8 | 32.0 | 33.2 | +1.2 |
| BCI IV-2b‡ | Freq. | 67.9 | 63.9 | 71.9 | +8.0 | 69.1 | 65.4 | 75.9 | +10.5 |
| | Wave. | 63.1 | 71.0 | 70.4 | −0.6 | 64.4 | 70.8 | 73.8 | +3.0 |
| | Dyn. | 71.7 | 72.3 | 78.6 | +6.3 | 73.5 | 72.5 | 79.4 | +6.9 |
| BNCI2015001‡ | Wave. | 55.1 | 58.9 | 57.1 | −1.8 | 55.3 | 58.9 | 60.3 | +1.4 |
| | Dyn. | 56.6 | 59.0 | 63.5 | +4.5 | 64.1 | 60.2 | 68.2 | +8.1 |
| | Proj. | 62.9 | 60.9 | 60.2 | −0.7 | 64.8 | 63.5 | 64.5 | +1.0 |
| BNCI2014009‡ | Freq. | 52.4 | 53.4 | 53.5 | +0.1 | 51.2 | 51.8 | 54.4 | +2.6 |
| | Wave. | 73.6 | 74.3 | 70.3 | −4.1 | 76.2 | 75.7 | 74.3 | −1.4 |
| | Rel. | 53.7 | 54.0 | 53.2 | −0.8 | 55.1 | 55.1 | 54.2 | −0.9 |
| | Dyn. | 65.3 | 73.5 | 72.7 | −0.8 | 67.9 | 75.3 | 74.9 | −0.4 |
| | Proj. | 53.3 | 53.6 | 55.4 | +1.8 | 54.5 | 54.6 | 56.4 | +1.9 |
| Nakanishi2015‡ | Freq. | 84.2 | 80.0 | 82.5 | +2.5 | 84.2 | 82.1 | 83.5 | +1.4 |
| | Wave. | 38.0 | 86.5 | 84.4 | −2.1 | 33.5 | 87.4 | 86.9 | −0.6 |
| | Rel. | 15.0 | 12.5 | 12.7 | +0.2 | 17.2 | 12.9 | 15.1 | +2.2 |
| | Dyn. | 32.8 | 87.5 | 85.0 | −2.5 | 27.2 | 88.6 | 87.9 | −0.7 |
| | Proj. | 8.3 | 8.8 | 8.4 | −0.4 | 6.6 | 7.2 | 9.4 | +2.2 |
| EEGMat | Freq. | 70.6 | 69.8 | 73.1 | +3.3 | 82.9 | 83.7 | 85.4 | +1.7 |
| | Wave. | 65.5 | 63.5 | 64.7 | +1.2 | 70.5 | 67.2 | 75.7 | +8.5 |
| | Rel. | 70.9 | 71.4 | 72.0 | +0.6 | 81.5 | 82.5 | 79.8 | −2.7 |
| | Dyn. | 67.7 | 61.9 | 65.5 | +3.5 | 76.0 | 66.8 | 78.6 | +11.9 |
| | Proj. | 62.5 | 59.5 | 63.4 | +3.9 | 67.4 | 59.1 | 65.6 | +6.5 |

## G COMPLETE RESULTS

This appendix gives the results behind Table 2: Tables 18–21 give every view on every dataset and setting beside the strongest published foundation model. Columns are the pretrained encoder of each view unless stated. "n/a" marks a view whose operator is undefined for the dataset (three electrodes on BCI IV-2b; a five-second support on BNCI2015001 for the Frequency and Relation views), and "–" a run not made. *CHB-MIT inputs contain 27,436 windows, against 29,840 in the benchmark. [p]Features reduced by a fixed count-sketch projection to 4,096 columns. Freq., Wave., Rel., Dyn. and Proj. are the Frequency, Waveform, Relation, Dynamics and Projective views.

**Which view leads.** Without alignment, the Dynamics view is the best single view in 22 of the 36 dataset–setting cells of Table 2, including ten of the twelve motor-imagery cells, and the Waveform view in four. The Frequency view leads frozen across subjects on the four tasks defined by spectral power: steady-state responses, emotion, seizure detection and mental arithmetic. The Relation view, which reads coupling between sensors, leads within subject on mental arithmetic and, after fine-tuning, on four-class motor imagery, where spatial covariance is the classical feature (Barachant et al., 2012). The Projective view leads after fine-tuning within subject on BNCI2015001. The largest margins over the Dynamics view are within subject, 20.0 points on mental arithmetic and 13.0 on four-class motor imagery. On both P300 datasets the Frequency, Relation and Projective views stay within about six points of chance in every setting in which they were run, and the best view is always the Waveform or the Dynamics view. The Waveform view also keeps each sensor's phase, and it leads within subject on steady-state responses, whose phase is locked to the stimulus.

**Comparison with published models.** Without alignment, a single view also ties the best published model in one cell; its largest margins are 17.6 points within subject on mental arithmetic and 15.7 on

Table 18: **LOSO, frozen**, benchmark probe, without alignment (mean of three classifier seeds). Encoder features on BNCI2014008, EEGMat, SEED and CHB-MIT pass the count-sketch projection (1,024 or 4,096 columns). The recentred Relation operator of BCI IV-2a (48.43) and the Relation encoder of Nakanishi2015 (11.89), normalised with statistics that include the held-out subject, are transductive and excluded (excl.). The Waveform and Dynamics columns use the 160,000-update encoders.

| Dataset | Freq. | Wave. | Rel. | Dyn. | Proj. | Best operator | Best FM |
|---|---|---|---|---|---|---|---|
| BCI IV-2a | 29.60 | 33.63 | 35.60 | 40.90 | 26.44 | 36.97 | 48.24 (Neuro-GPT) |
| BCI IV-2b‡ | 70.54 | 72.21 | n/a | 76.25 | n/a | 72.03 | 75.57 (Neuro-GPT) |
| BNCI2015001‡ | n/a | 57.38 | n/a | 65.40 | 59.64 | 59.00 | 63.43 (BIOT-6D) |
| BNCI2014008‡ | 50.00 | 50.08 | 50.00 | 50.42 | 50.00 | 53.18 | 67.11 (BENDR) |
| BNCI2014009‡ | 50.10 | 63.77 | 50.00 | 69.60 | 50.00 | 69.15 | 74.13 (EEGMamba) |
| Nakanishi2015‡ | 80.62 | 74.16 | excl. | 72.80 | 9.07 | 81.01 | 85.47 (EEGPT) |
| EEGMat | 72.62 | 69.88 | 72.07 | 70.28 | 62.90 | 72.13 | 68.58 (Neuro-GPT) |
| SEED‡ | 49.46 | 48.07 | 38.75 | 47.78 | n/a | 45.70 | 52.46‡ (LaBraM) |
| CHB-MIT*‡ | 69.92 | 58.40 | n/a | 66.36 | n/a | 63.28 | 75.21 (CBraMod) |

Table 19: **LOSO, full fine-tuning** under the benchmark's protocol.

| Dataset | Freq. | Wave. | Rel. | Dyn. | Proj. | Best FM |
|---|---|---|---|---|---|---|
| BCI IV-2a | 28.74 | 34.14 | 34.88 | 46.03 | 25.81 | 53.03 (CBraMod) |
| BCI IV-2b‡ | 71.57 | 71.94 | n/a | 75.60 | n/a | 77.70 (Neuro-GPT) |
| BNCI2015001‡ | n/a | 56.04 | n/a | 63.38 | 60.12 | 64.14 (LaBraM) |
| BNCI2014008‡ | – | 63.22 | – | 61.83 | – | 69.91 (CBraMod) |
| BNCI2014009‡ | 51.10 | 71.13 | 51.75 | 70.86 | 56.04 | 77.30 (CBraMod) |
| Nakanishi2015‡ | 83.85 | 76.50 | 14.70 | 91.01 | 9.80 | 92.94 (BENDR) |
| EEGMat | 73.61 | 67.22 | – | 70.09 | 63.52 | 72.62 (Neuro-GPT) |
| SEED‡ | – | 49.83 | – | 51.69 | n/a | 53.61 (CBraMod) |
| CHB-MIT*‡ | 68.08 | 66.45 | n/a | 71.19 | n/a | 78.12 (LUNA-Base) |

Table 20: **Within subject, linear probe** under the benchmark's protocol. On BNCI2014008 and BNCI2014009 the Waveform and Dynamics columns use the 160,000-update encoders.

| Dataset | Freq. | Wave. | Rel. | Dyn. | Proj. | Best FM |
|---|---|---|---|---|---|---|
| BCI IV-2a | 32.92 | 30.43 | 37.33 | 38.04 | 30.17 | 49.82 (Neuro-GPT) |
| BCI IV-2b‡ | 74.54 | 70.69 | n/a | 76.93 | n/a | 76.69 (Neuro-GPT) |
| BNCI2015001‡ | n/a | 54.29 | n/a | 61.59 | 61.51 | 67.80 (BIOT-6D) |
| BNCI2014008‡ | 51.48 | 55.51 | 51.40 | 58.93 | 52.47 | 61.45 (EEGMamba) |
| BNCI2014009‡ | 52.56 | 60.51 | 50.15 | 62.45 | 52.03 | 65.79 (EEGMamba) |
| Nakanishi2015‡ | 62.35 | 93.93 | 17.80 | 85.49 | 10.80 | 78.19 (BIOT-6D) |
| EEGMat | 86.34 | 68.83 | 94.68 | 79.86 | 83.18 | 77.08 (BIOT-6D) |
| SEED‡ | 55.90$^{p}$ | 52.34 | 40.93 | 57.24 | n/a | 52.32 (CBraMod) |
| CHB-MIT‡ | 72.22 | 76.40 | n/a | 77.67 | n/a | 87.77 (CBraMod) |

Table 21: **Within subject, full fine-tuning** (mean of three seeds).

| Dataset | Freq. | Wave. | Rel. | Dyn. | Proj. | Best FM |
|---|---|---|---|---|---|---|
| BCI IV-2a | 34.48 | 31.10 | 48.80 | 35.84 | 27.67 | 50.34 (CBraMod) |
| BCI IV-2b‡ | 74.24 | 63.79 | n/a | 76.50 | n/a | 77.39 (CBraMod) |
| BNCI2015001‡ | n/a | 55.75 | n/a | 60.54 | 65.54 | 70.30 (CBraMod) |
| BNCI2014008‡ | 50.28 | 53.45 | 51.91 | 55.13 | 52.39 | 61.61 (Neuro-GPT) |
| BNCI2014009‡ | 50.86 | 57.38 | 51.83 | 58.99 | 52.22 | 65.50 (EEGMamba) |
| Nakanishi2015‡ | 83.95 | 94.75 | – | 90.64 | 12.65 | 85.91 (BENDR) |
| EEGMat | 87.11 | 70.14 | 95.60 | 75.62 | 70.37 | 86.11 (BIOT-6D) |
| SEED‡ | 62.41 | 59.75 | 44.66 | 62.73 | n/a | 55.90 (Neuro-GPT) |
| CHB-MIT‡ | 84.61 | 72.81 | n/a | 83.60 | n/a | 88.54 (CBraMod) |

steady-state responses. With alignment, the frozen cross-subject multi-view results rise from 47.7 to 60.3 on SEED and from 70.2 to 77.1 on CHB-MIT. The View column of Table 2 takes the best of the five views in each cell, as FM takes the best of twelve published models; Section 5.4 reports how much of it a choice without test labels retains.

**EEGMat, all views and both controls.** EEGMat is the one dataset on which every view was scored frozen with both controls, with and without alignment; Section 5.3 draws on it (Table 22).

Table 22: **EEGMat, every view with both controls** (36-fold LOSO, benchmark probe, classifier seed 0, native width).

| | Without alignment | | | With alignment (transductive) | | |
|---|---|---|---|---|---|---|
| View | Operator | Random | Pretrained | Operator | Random | Pretrained |
| Frequency | 71.7 | 68.2 | 73.1 | 81.9 | 79.4 | 84.6 |
| Relation | 69.5 | 65.2 | 72.2 | 79.8 | 79.0 | 78.9 |
| Dynamics | 66.9 | 59.0 | 68.0 | 75.9 | 65.6 | 76.9 |
| Waveform | 61.3 | 58.7 | 63.3 | 66.0 | 66.9 | 72.2 |
| Projective | 53.6 | 54.1 | 62.4 | 66.8 | 59.5 | 66.1 |

## H WHERE THE LEARNING GAIN APPEARS

This appendix examines why the learning gains of Section 5.3 differ across views and tasks: which structures pretraining strengthens, what it adds to subject identity, and why the Waveform view loses on P300 under the benchmark's probe.

**Class share.** The test of Section 5.3 fixed its bar in advance: sign agreement in at least 13 of 16 view–dataset pairs and a Spearman correlation of at least 0.6 between the within-subject learning gain and the ratio, pretrained over random, of the share of feature variance explained by class. The class-share ratio meets it, with 14 of 16 and $\rho = 0.78$ (Figure 9), while the ratio of the variance explained by subject does not predict the gain ($\rho = 0.14$; Table 23). The two pairs whose signs disagree have gains below one point. The test establishes an association, and reading it as tracking how predictable the discriminative structure is from context is our interpretation. For the Waveform view the class share rises by a factor of 1.37 on steady-state responses and 1.11 on mental arithmetic, and falls to 0.79 on motor imagery and 0.49 on P300, where the deflection stays decodable (Appendix H.1): pretraining reallocates variance and keeps the signal.

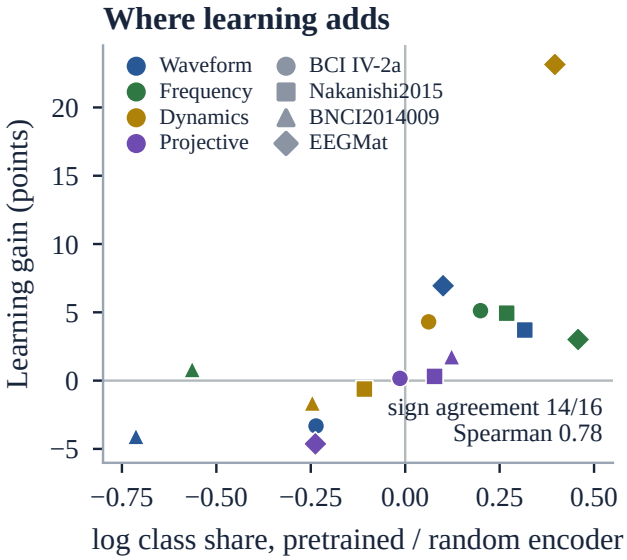



Figure 9: **The learning gain follows the class share** (closed-form ridge readout).

**Test details.** The test uses the four datasets with within-subject pairs of pretrained and random encoders for four views (BCI IV-2a, Nakanishi2015, BNCI2014009, EEGMat). The sign of the within-subject learning gain should agree with the sign of $\log(c_{\text{pre}}/c_{\text{rand}})$ in at least 13 of 16 pairs, with a Spearman correlation of at least 0.6, where $c$ is the share of feature variance explained by class means within subject after label-free global column standardisation. The gain is measured with a class-balanced closed-form ridge fitted on each subject's calibration trials, its penalty chosen by stratified three-fold cross-validation inside them.

Table 23: **The class-share ratio predicts the learning gain, the subject-share ratio does not** (sign agreement 14/16; Spearman 0.78 and 0.14).

| Dataset | View | Learning gain | Class share ratio | Subject share ratio |
|---|---|---|---|---|
| BCI IV-2a | Waveform | $-3.32$ | 0.79 | 1.49 |
| | Dynamics | $+4.30$ | 1.06 | 5.97 |
| | Frequency | $+5.12$ | 1.22 | 1.51 |
| | Projective | $+0.16$ | 0.99 | 0.57 |
| Nakanishi2015‡ | Waveform | $+3.70$ | 1.37 | 1.21 |
| | Dynamics | $-0.62$ | 0.90 | 3.94 |
| | Frequency | $+4.94$ | 1.31 | 2.02 |
| | Projective | $+0.31$ | 1.08 | 0.66 |
| BNCI2014009‡ | Waveform | $-4.08$ | 0.49 | 2.78 |
| | Dynamics | $-1.64$ | 0.78 | 4.12 |
| | Frequency | $+0.82$ | 0.57 | 4.03 |
| | Projective | $+1.75$ | 1.13 | 0.68 |
| EEGMat | Waveform | $+6.94$ | 1.11 | 1.39 |
| | Dynamics | $+23.15$ | 1.49 | 3.09 |
| | Frequency | $+3.01$ | 1.58 | 2.40 |
| | Projective | $-4.63$ | 0.79 | 0.62 |

**Subject identity.** Across subjects a second effect appears. Relative to the random encoder, the pretrained Waveform, Frequency and Dynamics features devote 1.2 to 6 times more variance to subject identity, which a cross-subject readout must ignore. Across the frozen single feature sets read by the same closed-form ridge under leave-one-subject-out, the mean learning gain is $+0.82$ points without alignment and $+2.52$ with it, on the same folds. Alignment removes each subject's mean shift, the part of the variance that the pretrained features devote to subject identity, which explains the gap between the within-subject and cross-subject results on emotion and seizure detection.

### H.1 Readout budget on the Waveform P300 features

The three analyses below locate the P300 deficit of the pretrained Waveform encoder (Section 5.3) in the readout (Hewitt & Liang, 2019; Kumar et al., 2022a).

**The deflection remains decodable.** We first checked whether pretraining removed the P300 from the features. A class-balanced closed-form ridge at a fixed penalty ($a = 1$) scores the operator, the random encoder and the pretrained encoder at 74.81, 75.88 and 73.70 over ten leave-one-subject-out folds; the best value over the penalty grid is an upper bound for each control, printed for reference (Table 24). A ridge decoder from the features back to the held-out subject's one-second window recovers the Target-minus-non-Target difference wave with slope 0.966, 0.923 and 0.903, and 0.986, 0.969 and 0.958 at Cz and Pz between 0.2 and 0.6 s. Band by band the pretrained features reconstruct the window better than the random ones, 0.945 against 0.926 below 4 Hz, where the P300 lies, and 0.706 against 0.505 above 30 Hz (Figure 10). The P300 is thus still in the features.

**The geometry changes.** What pretraining changes is the geometry that the readout faces. The pretrained features are more isotropic: effective rank 56.1 against 34.6, with 30% of the variance in the first ten principal directions against 45%. The discriminant therefore lies in lower-variance directions, and a readout that shrinks more loses more: from its best penalty to the largest on the grid, the pretrained encoder gives up 4.30 points and the random encoder 1.72.

**The benchmark's probe is a strongly shrinking readout.** The benchmark's probe shrinks strongly, which explains the deficit. Varying only the probe's epochs and dropout, with features, folds, optimiser

and learning rate fixed, gives at classifier seed 0: the random encoder reads 69.16 at the benchmark's budget (20 epochs, dropout 0.5), 69.70 at 100 epochs, 71.67 at 20 epochs without dropout and 71.84 at 100 epochs without dropout; the pretrained encoder reads 62.98, 65.54, 66.53 and 69.92. Relaxing the readout closes the gap between the two from 6.18 to 1.93 points, close to the 2.2 points that separate them under the closed-form ridge at a fixed penalty. The results tables use the benchmark's budget alone, because only it is comparable with published rows.

Table 24: **Under a ridge readout the three Waveform feature sets read the P300 about equally**, and the pretrained features are more isotropic (BNCI2014009, ten folds, without alignment). ERP slope: decoded against true difference wave; effective rank: $(\sum \lambda_i)^2 / \sum \lambda_i^2$.

| Features | Width | $a = 0.1$ | $a = 1$ | $a = 10$ | Best ($a$) | ERP slope | Eff. rank |
|---|---|---|---|---|---|---|---|
| Operator | 8,192 | 73.67 | 74.81 | 73.47 | 74.81 (1.0) | 0.966 | 29.3 |
| Random encoder | 11,264 | 74.92 | 75.88 | 74.25 | 75.97 (0.3) | 0.923 | 34.6 |
| Pretrained encoder | 11,264 | 74.66 | 73.70 | 70.35 | 74.66 (0.1) | 0.903 | 56.1 |

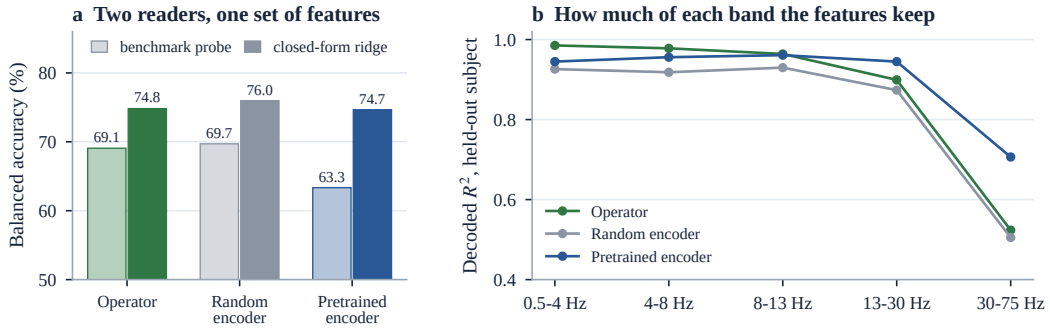



Figure 10: **The same Waveform features lose under the benchmark's probe and hold under a ridge readout** (a), and the pretrained features reconstruct the P300 band below 4 Hz better than the random ones (b).

# I Further related work

This appendix gives the lineage that Section 2 summarises.

**Classical EEG features.** The operators of OpView descend from classical EEG analysis, which has produced time-locked averages (Luck, 2014), band power (Pfurtscheller & Lopes da Silva, 1999; Welch, 1967), common spatial patterns (Blankertz et al., 2008; Ang et al., 2012), Riemannian covariance (Barachant et al., 2012; Congedo et al., 2017), coherency (Nolte et al., 2004), canonical correlation for steady-state responses (Lin et al., 2006; Nakanishi et al., 2018), periodic and aperiodic spectra (Donoghue et al., 2020), polarity-free microstates (Michel & Koenig, 2018) and Euclidean alignment (He & Wu, 2020).

**Spectral inputs and targets in EEG foundation models.** Several released models already read spectral information: BIOT tokenises short-time spectra (Yang et al., 2023), LUNA adds Fourier magnitude and phase to temporal patches (Döner et al., 2025), CBraMod embeds each patch through its spectrum as well as its samples (Wang et al., 2025), and LaBraM trains its tokenizer to predict Fourier amplitude and phase (Jiang et al., 2024). Each trains one representation with one objective; our views choose the observation space, its invariances and the objective together, one view at a time.

**Preconditioning.** In optimisation, a preconditioner is a fixed change of variables that makes a solution easier to reach (Nocedal & Wright, 2006). Deep learning applies the idea to inputs (LeCun et al., 1998), activations (Ioffe & Szegedy, 2015) and updates (Amari, 1998; Martens & Grosse, 2015; Kingma & Ba, 2015), and input representations set how fast functions are learned (Jacot et al., 2018; Arora et al., 2019; Rahaman et al., 2019; Tancik et al., 2020).

**Measuring representations.** Usable information separates structure present in data from structure a bounded learner can use (Xu et al., 2020); probes measure what a frozen representation offers a readout (Alain & Bengio, 2017; Hewitt & Liang, 2019; Kumar et al., 2022a); and audits of EEG foundation models separate what a representation encodes from what a task uses (Tang et al., 2026; Bindra & Panwar, 2026).