Imperial College London

MSc Individual Project

Imperial College London

Department of Computing

Evaluating Computational Brain Models as Dimensionality Reduction Methods for EEG

Author:
Lorenz Heiler

Supervisor:
Dr. Pedro Mediano
Dr. Gregory Scott

Second Marker:
Dr. Antoine Cully

September 10, 2025

Abstract

Electroencephalography (EEG) is high-dimensional and noisy, which makes compact representations essential for analysis and interpretation. Data-driven approaches such as principal component analysis, curated time-series features[1], and autoencoders often excel at prediction but offer limited physiological meaning. Computational brain models (CBMs) can invert biophysical simulators to yield parameter vectors with explicit neural semantics. This thesis benchmarks CBMs as dimensionality-reduction methods against strong statistical baselines within a single pipeline built on the Temple University Hospital Abnormal EEG Corpus (TUH-AB), a public clinical dataset curated for normal-versus-abnormal screening with subject-disjoint splits (version 3.0.1, 2,993 sessions).

We evaluate four CBM families (cortico–thalamic, Jansen–Rit, Wong–Wang, and Hopf) alongside PCA on power spectra, a spectral autoencoder, an EEGNet-style autoencoder, and the catch22 feature set. In addition, we provide an amortised parameter-inference implementation for the cortico–thalamic model as a proof of concept for hybrid approaches. We assess downstream classification performance on abnormal versus normal EEG and on biological sex. In addition, we evaluate information content via mutual information with the task labels, the geometry and cluster quality of the latent space, and dimensionality efficiency, defined as the fraction of latent dimensions actively encoding EEG.

On averaged spectra with small latent spaces, CBMs are compact and interpretable, with the cortico–thalamic model using amortised inference reaching 74.1% abnormal-screening accuracy. This is close to PCA on spectra at 73.5% and to a spectral autoencoder at 74.4%. The best sex accuracies in this setting are 55–60%. With per-channel inputs and larger latent spaces, performance rises across the board. The catch22 feature set reaches 79.3% abnormal accuracy and 62.0% for sex, while the cortico–thalamic model with amortised inference attains 78.4% abnormal accuracy and remains competitive with the spectral autoencoder at 78.2% and an EEGNet autoencoder at 78.3%.

Although CBMs are slightly below commonly used data-driven approaches on these tasks, latent-quality analyses show that they use capacity efficiently and provide mostly stable geometry. Learned and handcrafted features encode more class-relevant information on average. A hybrid design that combines a mechanistic forward model with learned parameter inference narrows this gap and often matches the strongest baselines we implement, while preserving physiological interpretability. Overall, CBMs are viable dimensionality-reduction tools that complement data-driven methods by offering biophysically grounded coordinates and a clear route to hybrid designs that balance interpretability and task utility.

Acknowledgements

I would like to express my sincere gratitude to my supervisors, Dr. Pedro Mediano and Dr. Gregory Scott, for their guidance, encouragement, and valuable feedback throughout the course of this thesis. I greatly enjoyed working with them, as they fostered a professional yet welcoming environment in which I always felt comfortable asking questions and exploring ideas. I am thankful to my fellow students and lab colleagues for their insights, helpful discussions, and for creating a collaborative and motivating environment. I also appreciate the support from my university in providing the necessary resources and facilities for this research. Finally, I am deeply grateful to my family and friends for their understanding, patience, and continuous support throughout my studies. In particular, I thank Alison and Kerim for spending every day in the library with me, and my family back in Germany for their encouragement and support from afar.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

A central aim of computational neuroscience is to understand how large-scale patterns of brain activity emerge in scalp recordings [4]. Electroencephalography (EEG) provides a noninvasive and millisecond-precise view of this activity, yet the signals are high dimensional and strongly affected by differences between people, recording sessions, and electrode setups. To make sense of this complexity, we seek *dimensionality reduction*: representing the data in a smaller set of coordinates that preserve information relevant for physiology and clinical use while filtering out nuisance variability. Done well, such representations make patterns easier to see, compare, and interpret, and they open a path toward linking scalp signals back to the neural mechanisms that generate them.

Two broad strategies are used to construct such representations. The first is *data-driven*, optimising statistical embeddings without explicit reference to underlying physiology. These methods are designed to achieve task-specific objectives, such as variance maximisation or independence of components, and often perform very well in supervised settings [5, 6, 7, 8]. However, their latent representations are typically abstract and not mechanistically interpretable, which makes post hoc analysis challenging and sometimes fragile.

Another approach in computational neuroscience is to use *computational brain models*, which typically formalise neural population dynamics through systems of differential equations. These models are governed by a limited set of parameters with biophysical priors and interpretable bounds. By tuning the parameters to minimise the mismatch between simulated and observed EEG signals, one obtains a compact representation whose coordinates are physiologically meaningful. Despite their differences, both data-driven and mechanistic strategies share the same goal: to compress high-dimensional EEG into a lower-dimensional space that preserves information of interest while filtering out nuisance variability.

Concrete examples help illustrate this distinction. Data-driven approaches span curated time-series descriptors (e.g., `catch22` [1]), linear embeddings such as PCA [9] applied to power spectral densities (PSDs), and neural encoders like autoencoders or compact convolutional networks in the style of EEGNet [10]. On the mechanistic side, we focus on three representative classes of computational brain models commonly applied to EEG and MEG: neural mass and neural field models (e.g., Jansen–Rit [11], cortico-thalamic variants [12]), dynamic mean-field nodes (e.g., Wong–Wang [13]), and normal-form oscillators (e.g., Hopf/Stuart–Landau [14]). These models describe interacting populations with a small set of state variables and interpretable parameters such as synaptic gains, time constants, coupling strengths, and delays. Fitting them to *observed EEG*, in this work primarily power spectra, yields parameter vectors that function as *mechanistic* coordinates. This promises parsimony and interpretability, while raising questions about expressivity, identifiability, and robustness on clinical data. Yet, while CBMs are often used as a form of dimensionality reduction, they have rarely been compared directly to data-driven approaches under the same conditions [4, 15, 16].

Aim. This thesis asks whether parameters extracted from fitting computational brain models can serve as useful low-dimensional embeddings for clinical EEG, and how they compare with common data-driven baselines under the same conditions. We build a unified pipeline on the Temple University Hospital Abnormal EEG Corpus (TUH-AB)[17, 18], a public clinical dataset curated for normal-versus-abnormal screening, and evaluate both families of methods on the same inputs and subject-disjoint splits. We also design a hybrid feature-extraction method that combines a

mechanistic forward model with learned parameter inference[19], offered as a proof of concept to examine the potential and limitations of hybrid approaches. This makes us focus on four questions that together define the evaluation:

1. *Sufficiency for downstream tasks.* Do mechanistic latents preserve information required for clinical endpoints (abnormal vs. normal and sex) at a level comparable to data-driven embeddings?

2. *Information content and geometry.* How do the two families differ in compactness, redundancy, and the preservation of neighbourhood structure from input to latent space?

3. *Interpretability and latent structure.* How do CBM and learned embeddings compare in interpretability and in latent structure quality (efficiency, independence, and geometry preservation)?

4. *Hybridisation.* Can we blend mechanistic priors with learned inference to recover discriminative structure without sacrificing interpretability?

With these questions guiding our evaluation, we present the following main contributions of this work.

Contributions and outline.

1. **Unified and modular benchmark.** A single TUH-AB–based pipeline with standardised preprocessing, inputs, splits, and metrics, designed for easy extension to new methods and datasets.

2. **Unbiased comparison.** Side-by-side evaluation of cortico-thalamic, Jansen-Rit, Wong-Wang, and Hopf models against curated baselines: `catch22`, PCA on PSDs, a PSD autoencoder, and an EEGNet autoencoder.

3. **Mechanistic embeddings.** Evidence that parameter-based embeddings are compact and interpretable, with utility that improves with capacity or spatial granularity.

4. **Hybrid proof of concept.** A single amortised-inference variant for the cortico-thalamic model that recovers much of the discriminative structure of statistical encoders while keeping physiological grounding.

The thesis proceeds as follows: Chapter 2 reviews background on EEG representations and computational brain models. Chapter 3 describes our methodology including the dataset, preprocessing, model families, and evaluation. Chapter 4 reports results and ablations, including the hybrid proof-of-concept. Chapter 5 discusses implications, limitations, and future work.

Chapter 2

Background

2.1 Why Dimensionality Reduction for EEG?

Electroencephalography (EEG) produces high–dimensional signals that mix neural and nuisance variability. A single segment can be represented as a tensor $x \in \mathbb{R}^{C\times T}$ (channels $\times$ time), yet its structure differs across subjects, sessions, montages, and hardware. Typical sample sizes are modest relative to this dimensionality, so a central goal is to construct compact representations $\phi(x) \in \mathbb{R}^d$ with $d \ll CT$. Such representations should preserve information relevant to clinical and physiological questions while being robust to nuisance factors. Formally, we seek embeddings that are *sufficient* for downstream tasks and *stable* across common perturbations, ideally satisfying invariances such as $\phi(g \cdot x) \approx \phi(x)$ for transformations g encoding referencing changes or small temporal shifts.

Design Requirements. An EEG representation must meet several requirements. These criteria guide our method choices in Sections 2.2–2.3 and shape the design of our evaluations:

- *Compactness and sufficiency:* low-dimensional (d small) while retaining task-relevant information.
- *Stability:* low sensitivity to mild artefacts, and session/hardware variations, and also predictable behaviour under re-sampling and windowing.
- *Physiological interpretability:* coordinates that can be related to spectral bands, topographies, or biophysical parameters.
- *Geometry preservation:* local neighbourhoods and global structure in the original feature space (e.g., PSD manifolds) are reflected in latent space to avoid spurious clusters.
- *Generalisation and transfer:* robust performance under subject shift and label scarcity, as well as compatibility with simple linear readouts to reduce overfitting.
- *Computational tractability and reproducibility:* feasible at corpus scale with clear hyperparameters and train/eval separation to avoid information leakage.

As outlined in the Introduction 1, this thesis studies *two different* paradigms side-by-side: mechanistic parameters as low-dimensional coordinates with explicit semantics, and data-driven embeddings.

Data-driven embeddings. These rely on statistical criteria to compress the data, for example hand-crafted time-series descriptors, linear spectral embeddings such as PCA on power spectral densities (PSDs), or learned encoders (autoencoders, EEG-specific CNNs, or more advanced models such as EEG2Rep or EEGDM) [9, 10, 20, 21, 22]. Their strength is flexibility and empirical performance, but interpretability is typically post hoc and sensitive to inductive biases.

Mechanistic (computational brain) models. These act as biophysical generative models whose parameters, when fitted to observed EEG, form the latent space (e.g., neural mass/field and mean-field models such as Jansen–Rit, cortico-thalamic, Wong–Wang, or Hopf/Stuart–Landau, as well as

whole-brain models like *The Virtual Brain*) [11, 12, 23, 13, 24, 25]. Their strength is physiological interpretability and parsimony, while challenges include potential non-identifiability and mismatch to empirical data.

In order to gain a better intuition for the two approaches we provide a quick comparison of some of their characteristics in table 2.1.

	Data-driven	**Mechanistic**
Primary inductive bias	Statistical (variance, reconstruction, invariances)	Biophysical (circuit motifs, transfer functions)
Latent semantics	Emergent / post hoc	Explicit (physiological parameters)
Data requirements	Often large unlabeled sets suffice	Less data but stronger priors; simulator access
Robustness to nuisance	Needs explicit handling (normalisation, augmentation)	Priors can help; mismatch can hurt
Identifiability	Not applicable in same sense	Central challenge (equifinality, degeneracy)
Computation	Training cost upfront; fast inference	Per-sample fitting or amortisation; may be costly
Typical outputs	Low-dim codes without units	Parameters with units/physiology

Table 2.1: Conceptual comparison of the two paradigms.

A further practical distinction concerns the data representation each approach operates on. Data-driven methods can infer compact latents either from raw waveforms or from spectral summaries[1, 10, 9], whereas mechanistic models are most often fitted in the frequency domain[2, 11, 26, 23]. Because many such models generate spectra directly, they allow straightforward comparison between simulated and observed activity without reconstructing full time series. This makes the *power spectral density* (PSD) a natural choice of representation in this work.

What are power spectral densities?
EEG signals are often studied in the frequency domain because brain rhythms are naturally organised into bands such as delta (1–4 Hz), theta (4–8 Hz), alpha (8–12 Hz), beta (13–30 Hz), and gamma (> 30 Hz)[27, 28]. These bands are linked to well-established physiological and clinical markers, for example alpha power in posterior regions during rest or beta activity in movement disorders[29, 30, 2]. The *power spectral density* (PSD) quantifies how the signal's power is distributed across frequencies. Formally, the PSD of a time series $x(t)$ is the Fourier transform of its autocorrelation function, and in practice it is estimated from finite data segments using methods such as Welch's periodogram [31]. PSDs reveal both narrow-band oscillatory peaks (e.g., alpha, beta) and the broadband $1/f$ background, and are relatively robust to phase variability and common artefacts. For these reasons, PSDs serve as a widely used and clinically meaningful representation of EEG.

For computational brain models, we restrict ourselves to power spectral densities (PSDs), since fitting CBMs to spectra is the standard and tractable approach, whereas direct fitting to raw, nonstationary EEG is often cumbersome due to noise and artefacts. Data-driven methods, by contrast, are not bound to the frequency domain. In this work we therefore evaluate both raw-segment approaches (e.g., `catch22` [1], EEGNet [10]) and spectral approaches (e.g., PCA, autoencoders applied to PSDs). Including the latter ensures that both model families can be compared on spectral inputs, while still allowing raw-time-series benchmarks on the statistical side.

In conclusion, data-driven and mechanistic embeddings represent two complementary strategies: the former emphasise statistical flexibility, while the latter impose physiological structure. To make their strengths and limitations tangible, the next section (2.2) surveys representative data-driven approaches, ranging from hand-crafted features to modern neural encoders.

Scope of the remainder of the chapter. This chapter is organised to provide the necessary background for the methods used in this thesis and to situate them within the current standards of the field. Section 2.2 surveys data-driven methods, from hand-crafted features to deep encoders. Section 2.3 introduces mechanistic models as latent spaces, spanning neural mass, cortico–thalamic, mean-field, and oscillator formulations. Section 2.4 then positions this work within the broader literature. Practical implementation details are deferred to Chapter 3.

2.2 Data-driven methods

Hand-crafted time-series feature approaches

A direct way to build EEG representations is to extract a predefined set of statistical features from short signal windows and use those values as the embedding. Two widely used libraries are *catch22* and *TSFRESH* [1, 32]. *catch22* offers a curated set of 22 features that capture distributional shape, linear and nonlinear autocorrelation, and temporal heterogeneity. *TSFRESH* exposes a much larger catalogue and then filters features using statistical tests, which in practice can produce on the order of 10^2–10^3 features per channel before selection. In EEG, these generic descriptors are often combined with domain features such as band powers in canonical frequency ranges and peak frequencies derived from power spectral density estimates[33]. The result is a vector of interpretable summary statistics per epoch that can be supplied to a classifier or clustering method.

The main advantage of hand-crafted features is interpretability. Each coordinate has a clear definition and is often linked to a physiological hypothesis, for example elevated frontal theta power or changes in signal entropy. This makes downstream analyses easier to explain. Feature extraction is also computationally light. Means, variances, spectra, and simple entropy measures are efficient to compute at scale, which enables rapid iteration and reproducible pipelines. Historically, such feature sets paired with conventional classifiers have produced strong baselines across many EEG tasks, especially when prior knowledge aligns with the target phenomena.

These benefits come with trade-offs. Hand-crafted features only work as well as the assumptions behind them. If the descriptors you choose do not cover the signal patterns that matter, the representation will miss or distort what is in the data. Very large, automatically generated feature sets can overfit and produce spurious correlations. Many descriptors measure nearly the same property, so careful selection or regularisation is needed to avoid learning from noise. Because each window is condensed into a few aggregates, fine temporal events can be washed out. Differences between subjects and sessions also shift absolute amplitudes and band powers. Without robust normalisation and harmonisation, models trained on these features may fail to transfer.

In practice, curated sets such as *catch22* help by reducing redundancy and keeping capacity low, while broader toolkits like *TSFRESH* are useful when combined with strict selection and robust scaling. Adding a small number of EEG-specific summaries, for example band powers or peak frequency, can cover spectral effects that generic descriptors may miss. Overall, hand-crafted features remain transparent, fast, and domain-aware. They work best with solid preprocessing and clear normalisation, and they serve as strong baselines and diagnostic tools. They are less suited to subtle or highly nonlinear structure, which motivates complementing them with learned encoders later in this chapter and contrasting both families with mechanistic representations in the following sections.

Principal Component Analysis (PCA)

PCA is a linear method that finds an orthogonal basis whose axes capture the largest variance in the data in descending order [9]. Truncating to the top K components yields a K-dimensional summary that preserves as much variance as possible for that dimensionality. In EEG, PCA is commonly applied to spectral summaries such as power spectral densities across frequencies and channels, or to channel-time features after basic preprocessing[34]. Fitting the projection on training data only and freezing it for evaluation avoids information leakage. The principal axes

can be inspected through their loadings, which often reveal familiar spectral patterns like broad $1/f$ structure or a dominant alpha peak[35].

PCA is attractive because it is unsupervised, fast, and often denoises by discarding directions with little variance. As a baseline it is transparent and easy to reproduce, and it provides compact features that work with simple linear readouts. Applied to PSDs it exploits the low intrinsic rank of scalp spectra and can be used either on channel-averaged inputs to emphasise global rhythms or on per-channel inputs to retain topographic information.

Its limitations are equally important. Maximising variance does not guarantee relevance for a task. High-variance directions can reflect confounds such as subject identity, amplifier differences, or artifacts, while clinically meaningful effects may live in lower-variance subspaces and be truncated. PCA is linear, so it cannot unfold nonlinear manifolds that arise from complex neurophysiology. The basis is data dependent, which reduces transfer across cohorts if preprocessing or recording conditions differ. Channel averaging improves stability but sacrifices spatial detail, whereas per-channel PCA preserves topography at the cost of higher dimensionality and potential overfitting in small datasets.

Deep neural network approaches

Deep models learn EEG representations directly from data by optimising flexible, multilayer function approximators. Unlike hand-crafted features or linear projections, they can capture nonlinear interactions across time, frequency, and space, and they can be trained without labels via reconstruction or self-supervised objectives. In practice, three design axes matter most: the *input view* (raw time series, power spectral density, or time–frequency maps), the *architecture* (convolutional encoders such as EEGNet[10], recurrent/temporal convolutions, or transformers), and the *objective* (reconstruction, contrastive agreement, masked prediction). Below we focus on approaches that are both widely used and relevant to our evaluation: spectral autoencoders, EEGNet-style spatio–temporal encoders (including autoencoder variants), and modern self-supervised pretraining.

Spectral and time–frequency autoencoders. Autoencoders (AEs) compress an input into a low-dimensional code and learn to reconstruct the input from that code [36, 37]. When the input is a per-window power spectral density (PSD) vector or a time–frequency map, the bottleneck is encouraged to retain the most salient spectral structure while discarding nuisance variation. Fully connected AEs on PSDs often recover latents aligned with broad spectral tone, alpha peak prominence, or $1/f$ slope. Convolutional AEs on spectrograms exploit local time–frequency patterns and can model co-modulation across bands. Compared with PCA on PSDs, nonlinear decoders capture curved manifolds and band interactions, offering better fidelity for the same latent dimensionality. The main limitation is that phase information and fine temporal dynamics within a window are lost. Reconstruction losses can also be dominated by broadband magnitude, under-representing narrow but clinically relevant peaks unless capacity and normalisation are chosen carefully [36, 37]. Variational formulations add stochasticity and regularise the code geometry but introduce additional optimisation hyperparameters [37]. In this thesis we use a PSD-AE as a canonical spectral learner to test how far a purely statistical nonlinear compressor can go before hitting the limits of the spectral view.

EEGNet and spatio–temporal encoders (and their AE variants). EEG-specific convolutional networks such as *EEGNet* encode domain priors directly into the architecture [10]. A first temporal convolution behaves like a learnable filterbank that approximates band-pass filters, a subsequent depthwise spatial convolution learns channel-wise topographies, a pointwise separable stage mixes features with few parameters[10]. This factorisation matches the structure of EEG, where informative content is organised jointly in frequency and scalp space. In supervised mode, EEGNet has been widely adopted across BCI and clinical tasks because it balances inductive bias, data efficiency, and compactness.

Repurposed as an *autoencoder* (EEGNet-AE), the same inductive biases can learn unsupervised latents from raw multi-channel windows: the encoder compresses time–channel segments to a code of tens of dimensions, and a mirrored decoder reconstructs the signal. Such models typically capture low-frequency components, oscillatory bursts, and canonical topographies[10]. First-layer temporal kernels resemble data-driven band-pass filters while depthwise spatial kernels align with interpretable maps (e.g., posterior alpha). Compared with spectral AEs, EEGNet-AEs retain tem-

poral locality and can represent waveform shape, onsets, and cross-channel timing. They are also compact enough to train on moderate datasets, helping generalisation across subjects. However, capacity must be controlled to avoid memorising session-specific artefacts. Architectural choices (kernel lengths, dilation, pooling) and normalisation strongly influence whether the code reflects physiology or confounds. Variational EEGNet variants further encourage smooth, disentangled latents but can blur fine temporal detail if the decoder is underpowered.

Self-supervised pretraining on unlabeled EEG. Self-supervised learning (SSL) pretrains an encoder on unlabeled EEG to produce task-agnostic embeddings that transfer under label scarcity and domain shift [20]. Three families dominate: *contrastive* methods create two views of the same segment with label-preserving augmentations and train the encoder so positives are close while others are pushed apart. *Asymmetric Siamese* methods align two augmented views without explicit negatives via an online–target architecture and *masked prediction*/predictive coding methods learn from context by reconstructing masked time regions or forecasting future representations. On EEG, augmentation is the crucial design choice because it encodes the invariances we want the representation to learn. Effective policies use small temporal shifts, mild time warping, band-limited perturbations, realistic additive noise, channel dropout, and light re-referencing jitter, all tuned to preserve neural content[38]. As a representative masked-prediction approach, *EEG2Rep* pretrains encoders on unlabeled EEG by reconstructing masked segments, yielding label-efficient, transferable embeddings [21]. When designed and tuned appropriately, self-supervised pretraining on unlabeled EEG *often* produces embeddings that are more invariant to subject/session variability than purely supervised training and that fine-tune effectively with limited labels [20, 39, 21, 40].

The main risks are that models rely on *spurious signals*, such as amplifier artifacts, electrode impedance effects, or subject-specific scaling, rather than neural dynamics. Negative-free objectives can *collapse* if the signal is weak or augmentations are too mild, meaning the encoder outputs nearly the same vector for everything. Contrastive objectives can be brittle if the batch is too small or the temperature is poorly tuned, because the model does not see enough informative negatives to learn meaningful boundaries. Augmentations can also go wrong: if they are too aggressive, they erase the very rhythms or transients we care about.

Strengths and limitations in practice. Deep encoders learn rich features that capture nonlinear coupling and joint spectro–spatial structure, and they often outperform hand-crafted and linear baselines when trained on diverse cohorts. Compact EEG-specific networks are comparatively data-efficient and somewhat interpretable: first-layer temporal filters behave like learned band-pass filters, and depthwise spatial kernels resemble scalp maps. The main limitations are reduced transparency relative to mechanistic parameterisations, sensitivity to dataset bias, and a tendency to encode confounds such as subject identity, montage, or site when normalisation is weak. Self-supervised pretraining reduces label demand and improves cross-subject stability, but it is sensitive to augmentation policy and can collapse if signals are weak. Autoencoders may allocate capacity to easy-to-reconstruct artefacts unless artefact-aware preprocessing or constraints are applied. The portability across cohorts relies on harmonised preprocessing and consistent scaling. Common pitfalls include embeddings dominated by artefacts or hardware signatures, reconstructions that smooth away informative narrow-band peaks, and SSL features that separate sessions rather than brain states.

For our thesis we use two representative deep approaches: a *spectral autoencoder* on PSDs and an *EEGNet-derived autoencoder* on raw multi-channel windows. Together they bracket the design space between a frequency-focused, stable summary and a spatio–temporal encoder with stronger inductive bias. Both are trained with subject-disjoint protocols and evaluated with simple downstream readouts. This pairing lets us quantify gains over PCA and compare learned latents with mechanistic parameter vectors in terms of compactness, stability, and physiological plausibility. Implementation details are deferred to Chapter 3.

Practical trade-offs and considerations

Generalisation. Linear baselines and curated feature sets generalise reliably when acquisition and preprocessing match, but they degrade under cross-site or cross-session shift. Deep encoders can learn subject-invariant structure if trained on sufficiently diverse data and regularised with

normalisation and augmentation. When shifts persist, simple domain-adaptation tactics help in practice: instance or channel normalisation to suppress scale cues, balanced batching across subjects, and subject-adversarial regularisers that discourage identity leakage. Throughout, we evaluate on subject-disjoint splits to reflect real deployment.

Interpretability. Hand-crafted features are directly explainable and easy to audit. PCA loadings and ICA scalp maps support partial interpretation. Deep latents are more opaque, but inspection of first-layer filters and spatial kernels, plus frequency-response analyses and input–gradient saliency, provide useful sanity checks.

Alignment with physiological priors and hybrids. Priors accelerate learning when correct but can constrain discovery. EEG-specific convolutions inject mild structure without hard-coding bands or sources[10].

In sum, we balance interpretability with flexibility by comparing three data-driven baselines—PCA, a spectral autoencoder, and an EEGNet-derived autoencoder—to mechanistic parameterisations, all evaluated on the same PSDs. Chapter 3 describes preprocessing and training. Chapter 4 reports compactness, stability, and downstream utility under subject-disjoint splits.

2.3 Mechanistic (Computational Brain) Models as Latents

In contrast to flexible data-driven methods, mechanistic approaches use explicit computational brain models as the source of an EEG representation. The core idea is to treat a biophysically grounded simulator of neural activity as a generative model for EEG and use its parameter vector as the latent representation when fitted [41]. This yields a low-dimensional coordinate system where each axis has a clear physiological meaning – often corresponding to properties like synaptic gains, time constants, or transmission delays [23]. The theoretical appeal is that these latent coordinates directly encode circuit-level features (e.g. excitation–inhibition balance or loop delays) instead of abstract statistical components. Such *mechanistic embeddings* promise interpretability and parsimony [23]: for example, in a classic neural mass model, one parameter may represent the excitatory synaptic strength and another the inhibitory time constant [11], grounding the representation in neurophysiology rather than arbitrary features.

Neural Mass and Neural Field Models

Neural mass models provide one of the most widely used frameworks for linking mesoscopic physiology to macroscopic EEG signals. The core idea is to lump the activity of thousands of neurons into a handful of aggregate variables, thereby describing a local cortical circuit with minimal but biophysically interpretable complexity. Instead of simulating individual spikes, these models track average firing rates and postsynaptic potentials of neuronal populations, which makes them computationally efficient while retaining physiological plausibility.

A canonical example is the Jansen-Rit model of a cortical column [11, 23]. It represents three interacting populations: pyramidal cells, excitatory interneurons, and inhibitory interneurons. Pyramidal cells provide the main output, while the two interneuron populations regulate excitation and inhibition. Synaptic input is transformed into postsynaptic potentials through kernels with characteristic time constants a and b, which correspond to excitatory and inhibitory synapses. A static sigmoid then maps mean membrane potential to firing rate. Key parameters of the model include the excitatory gain A, inhibitory gain B, and the synaptic rate constants. Adjusting these values alters the oscillatory behaviour of the model. For example, increasing the inhibitory gain tends to dampen or slow oscillations, while raising excitatory gain can push the system toward stronger rhythmicity. Within a certain parameter regime, the model naturally produces a resonant peak in the alpha band, making it a useful mechanistic account of cortical rhythms.

When fitted to empirical power spectral densities (PSDs), the Jansen-Rit model produces a set of latent coordinates: the gains, time constants, and connectivity parameters that best reproduce the observed EEG spectrum. Neural mass models define a *theory-driven latent space* where different EEG recordings can be compared in terms of underlying synaptic physiology rather than arbitrary signal features. At the same time, these models face challenges. Their parameter landscapes are

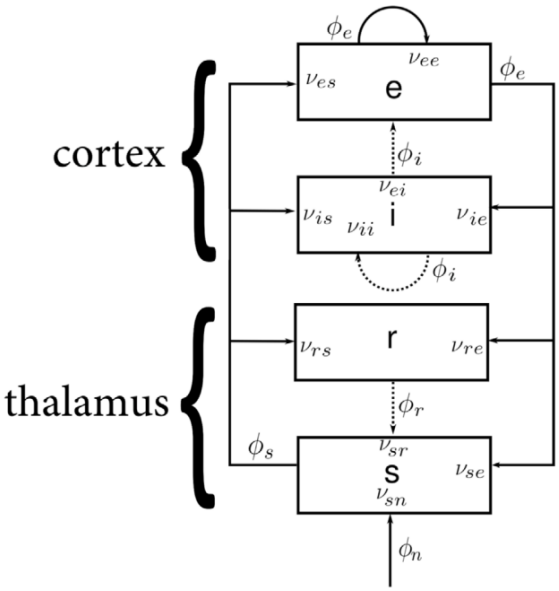


Figure 2.1: Structure of the corticothalamic system[2].

highly non-convex, different parameter sets can generate nearly indistinguishable spectra, and fitting them to noisy EEG data often requires robust optimisation strategies.

Cortico-thalamic models (CTMs) extend the neural mass and mean-field tradition by explicitly incorporating the recurrent loops between cortex and thalamus that play a central role in rhythm generation. These models typically include four interacting populations: cortical excitatory and inhibitory neurons, along with thalamic relay and reticular nuclei (see Figure 2.1) [42, 43, 2]. Long-range projections from cortex to thalamus and back introduce characteristic delays, while local synaptic interactions set distinct time constants. Together, these elements naturally reproduce key features of human EEG, most prominently the alpha rhythm.

Mathematically, CTMs are formulated as delay-differential equations. Around a stable resting state, they can be linearised, yielding frequency-domain transfer functions that map stochastic input to output spectra. The spectral shape depends on a small number of physiologically interpretable parameters, such as excitatory and inhibitory synaptic gains, synaptic time constants, and corticothalamic delays. By fitting these parameters to empirical EEG PSDs, one embeds each recording into a latent space that reflects thalamo-cortical loop properties rather than abstract signal statistics.

Relative to local cortical models like Jansen-Rit, CTMs provide a more global account of oscillations, emphasising how recurrent corticothalamic interactions shape spectral peaks and slopes. Unlike mean-field decision models such as Wong-Wang, which capture slow metastable dynamics, CTMs are tuned to stationary oscillatory structure and are particularly well suited for capturing variability in alpha-band activity. Their parameters offer a direct link to corticothalamic physiology, which is of both clinical and theoretical interest.

Inversion of CTMs can be achieved using population-based optimisers such as covariance matrix adaptation evolution strategies, or, as we also do it in this thesis, with modern amortised inference techniques where neural networks are trained to map spectra directly onto parameter estimates [19]. This allows both detailed subject-level fits and large-scale applications. The resulting parameter vectors serve as mechanistic embeddings that are interpretable by design. As with other models, however, identifiability issues and model mismatch remain important concerns, and parameter estimates must be treated with caution.

Taken together, neural mass and cortico-thalamic models show how biophysically grounded dynamical systems can act as latent representations of EEG data. They compress high-dimensional signals into low-dimensional coordinates that map onto circuit-level physiology. This makes them appealing for interpretability and hypothesis generation, while also highlighting the need for rig-

orous evaluation of robustness and utility.

Mean-field models (e.g. Wong–Wang)

Dynamic mean-field models extend the neural mass concept by incorporating nonlinear population dynamics often derived from simplified spiking networks. A prominent example is the Wong–Wang model, originally developed to describe decision-making circuits, which has become a useful node-level model in brain network studies[13, 44]. The Wong–Wang model captures the effective interaction of excitatory recurrence and inhibitory influence through a sigmoid input–output function, producing either stable fixed-point activity or oscillatory-like behavior depending on the parameters. In practice, this model can be tuned with parameters such as an *effective excitation strength* (recurrent self-excitation) and an external *input drive* current [13]. These parameters control the overall gain and stability of the population firing rate. When used as a latent embedding for EEG, the Wong–Wang model's fit reflects how excitable or inhibited a particular recording's underlying neural circuit might be. For instance, a higher fitted recurrent gain might indicate a circuit closer to a critical threshold of oscillation or persistent activity, whereas a higher input drive might reflect generally elevated firing levels. Like neural mass models, mean-field models yield a *biophysical parameter space* for representing data. Each EEG segment is encoded by the pair (or few) of parameters governing the mean-field equations that best reproduce the segment's PSD. This approach has the appeal of linking the latent features to properties like "net cortical excitatory tone" or "background input level," which align with theoretical accounts of EEG differences (e.g. sedation might manifest as reduced excitation drive in the model's fit).

Oscillator models (e.g. Hopf/Stuart–Landau)

Oscillator-based models provide a more abstract but mathematically tractable way to represent brain rhythms. A common choice is the normal form of a Hopf bifurcation (the Stuart–Landau oscillator), which serves as a minimal model for neural oscillations [45, 46]. In this formulation, a pair of differential equations (or a complex equation) governs a damped oscillator that can transition to sustained oscillations when a bifurcation parameter crosses a critical value. The *latent parameters* typically include the oscillator's natural frequency and a bifurcation coefficient that controls the qualitative regime (quiescent vs. oscillatory). By fitting a Hopf model to an EEG power spectrum, one essentially asks: *is there an intrinsic oscillatory mode that explains this spectrum, and what are its frequency and stability?* For example, a Hopf model might fit an EEG with a strong alpha peak by assigning a latent frequency near 10 Hz and a bifurcation parameter near the critical point (indicating the system is well-positioned to oscillate at alpha frequency). If the EEG is more broadband (no clear peak), the Hopf fit might stay in the subcritical regime (damped oscillator) with parameters that produce a $1/f$-like spectrum instead of a sharp tone [14]. Oscillator models thus embed each recording in a *space defined by oscillatory dynamics*: one coordinate encodes the dominant frequency and the other encodes how close the circuit is to self-sustained oscillations. This is a highly interpretable description, connecting directly to notions like "how strong is the endogenous rhythm?" and "at what frequency?".

Mapping Models to Spectra: Parameter Inversion

A unifying feature of mechanistic models is that each parameter set can be *mapped forward* into a predicted EEG spectrum. In many cases models have transfer functions or frequency-domain solutions that make this mapping explicit[2, 26]. For example, the Robinson cortico–thalamic neural field model yields an analytic expression for its power spectral density (PSD) as a function of synaptic gains and corticothalamic delays [43]. Similarly, linearised formulations of the Jansen–Rit neural mass provide closed-form transfer functions that link synaptic parameters to spectral output [23, 26]. When such analytic solutions are not available, one can simulate the full dynamical system and then convert the resulting time series into the frequency domain. In all cases, this forward mapping connects parameter vectors to spectral features in a transparent and mechanistic way.

This forward mapping enables *parameter inversion*: given an empirical EEG power spectrum, the task is to identify the model parameters that produce the closest match. Inversion effectively embeds each EEG recording into the parameter space of a chosen model family. The resulting

latent vector has physiological meaning by construction. For example, in a neural mass model the embedding might be expressed as $\theta = (A, B, a, b)$, corresponding to excitatory and inhibitory synaptic gains and their time constants. In a Hopf oscillator, the latent might reduce to a pair (f_0, γ), describing the central frequency and linewidth of an oscillatory mode. Each coordinate therefore represents a biophysically interpretable property of the circuit, and changes along these coordinates correspond to predictable alterations of the spectrum. An increase in inhibitory gain suppresses high-frequency power, while a shift in bifurcation parameter can toggle between noisy fluctuations and sustained oscillations.

Several approaches to inversion have been explored in the literature. Classical methods rely on *Bayesian inference*, such as the variational Laplace scheme used in Dynamic Causal Modelling (DCM), which estimates parameters and their uncertainty [47, 48]. *Population-based* strategies such as covariance matrix adaptation evolution strategy (CMA-ES) are gradient-free and handle noisy, multimodal objectives. They have been applied successfully to whole-brain and EEG model fitting and show stable performance in solver comparisons [49, 50]. When models are differentiable, *gradient-based optimisation* can also be applied. Recent work has demonstrated the feasibility of embedding neural mass models in machine-learning frameworks: for example, Momi et al. implemented a Jansen–Rit connectome model in PyTorch and used ADAM with automatic differentiation to fit subject-specific TMS–EEG responses [51]. A complementary trend is *amortised inference*, where regressors are trained on large synthetic datasets to approximate the inverse mapping directly. This regression-based strategy was first introduced in regression DCM for fMRI [19]. Such an approach shift the computational burden to an offline training phase, enabling near-instantaneous parameter estimation for new data in clinical-scale settings.

Challenges and Considerations

Despite their appeal, mechanistic latent spaces face important limitations. A foremost issue is *parameter non-identifiability*: many distinct parameter combinations can produce very similar spectra, a phenomenon often referred to as equifinality [52]. For example, changes in excitatory gain may be counterbalanced by adjustments in inhibitory time constants, leaving the alpha peak in the PSD essentially unchanged. As a result, the inverse mapping from spectra to parameters is typically ill-posed, and small amounts of noise or model misspecification can lead to large shifts in the inferred parameters.

Model mismatch is a second concern. All neural mass and field models are abstractions, and empirical EEG often contains features outside the model's representational scope. In such cases, the fitting process may force parameters into regimes that absorb the discrepancy, yielding values that no longer correspond directly to the intended physiology. This makes careful interpretation essential: a good numerical fit does not guarantee a correct mechanistic account.

Finally, practical issues arise from the computational demands of fitting. Unlike data-driven embeddings, which are often obtained in a single forward pass, mechanistic inversion requires solving an optimisation problem for every recording. This can be costly in large datasets, particularly when models are simulated iteratively or involve high-dimensional parameter spaces. Although recent approaches such as amortised inference or efficient black-box optimisers can mitigate this burden, the trade-off between accuracy, robustness, and scalability remains an open challenge.

2.4 Positioning within the Literature

Work on EEG dimensionality reduction (DR) and representation learning has largely developed along two separate tracks. On one side, **data-driven methods**, from classical approaches such as PCA and ICA to curated feature sets like catch22 and more recent deep neural encoders, aim to learn compact features directly from data. Comparative studies have systematically benchmarked these methods against one another in contexts such as emotion recognition or BCI decoding [53]. Such approaches can capture complex statistical regularities, yet the resulting latent factors are often abstract and difficult to interpret neurophysiologically. On the other side, **computational brain models (CBMs)** provide biophysically grounded descriptions of neural dynamics. CBMs

generate EEG/MEG rhythms from parameters with clear mechanistic meaning [41]. These models have primarily served as *forward generative tools*, used to simulate data or to test mechanistic hypotheses, but they are rarely employed as general-purpose DR methods because parameter inversion from empirical EEG is challenging [54]. As a result, the field has tended to either use statistical embeddings with little physiological grounding or mechanistic models with little connection to modern DR benchmarks.

A small but growing body of work has begun to explore integration between these traditions. The idea of *generative embedding*, mapping data into the parameter space of a biophysical model, was pioneered with fMRI: Brodersen and colleagues showed that DCM-derived parameters could classify patients more accurately than PCA features, while retaining clear biological interpretability [55]. Similar principles have since been applied to EEG, for example by fitting Jansen-Rit models to event-related EEG and using the inferred parameters as features for depression classification, where they outperformed conventional ERP features across classifiers [56]. Other extensions have used clustering in mechanistic parameter space to reveal subgroups of subjects [57], or synthetic EEG generated from neural mass models to test inversion algorithms and validate recovery of ground-truth parameters [58]. Recent work has also demonstrated that cortico-thalamic model parameters can track arousal states: Assadzadeh et al. showed that fitting a neural field model to EEG spectra yields parameters that cluster into wake, sleep, and brain-injured groups, with specific synaptic gains and delays differentiating states of consciousness [2]. More recently, hybrid approaches have emerged that blend physiological structure with machine learning, such as networks of trainable Hopf oscillators for sleep EEG reconstruction [59]. Together, these examples demonstrate the promise of treating CBM parameters as low-dimensional, interpretable embeddings. However, they remain isolated demonstrations, often tied to specific tasks or datasets, and they do not provide a general framework for comparing mechanistic and statistical approaches under common conditions.

What is notably absent from the literature is a systematic evaluation that places data-driven and mechanistic embeddings on equal footing. Comparative EEG studies typically restrict themselves to statistical methods [60, 61, 53], while CBM-based analyses validate mechanistic models in isolation or use them for specific hypothesis testing. No established benchmark exists where PCA, autoencoders, and neural mass or field models are all processed through the same pipeline and evaluated with the same metrics. Consequently, the field lacks a clear understanding of the trade-offs between the interpretability of mechanistic embeddings and the flexibility of purely statistical ones. Calls for integrative benchmarks in neuroscience [62] highlight this gap, emphasising the need to evaluate models not only on predictive accuracy but also on their alignment with brain mechanisms[15, 4]. Yet for EEG and MEG, such integrative evaluations have not been realized.

This thesis addresses this gap by unifying the two strands within a single framework. First, we treat mechanistic parameters as embeddings, reframing computational brain models not just as generative tools but as dimensionality reduction methods that map each EEG recording into a physiologically interpretable parameter space. Second, we introduce a method-agnostic benchmark that evaluates data-driven and mechanistic embeddings side by side, under standardised preprocessing, datasets, and evaluation metrics. This provides the first systematic comparison across families that were previously studied in isolation, allowing us to quantify the respective strengths and limitations of each. Third, we explore a hybrid amortised inference approach, in which a neural network is trained to predict CBM parameters directly from spectra. This combines the efficiency and robustness of deep learning with the interpretability of mechanistic latents, drawing on ideas from simulation-based inference [54, 19] but integrating them into an end-to-end DR framework.

In positioning CBMs alongside modern data-driven methods, our work makes explicit a trade-off that has so far remained implicit: interpretability versus flexibility. By embedding EEG data into mechanistic parameter spaces and comparing them directly with statistical embeddings, we provide a standardised evaluation that clarifies when mechanistic interpretability can be achieved without sacrificing performance, and when data-driven flexibility offers clear advantages. In doing so, this thesis responds to a clear need in the field and provides a new perspective on EEG dimensionality reduction: one that combines the *robustness* of machine learning with the *insight* of principled brain models.

Chapter 3

Methodology

3.1 Overview

This thesis investigates how different latent extraction paradigms encode clinically relevant EEG structure. We construct a unified pipeline that (i) standardizes clinical EEG (TUH-AB) via a rigorous preprocessing stack, (ii) derives latent representations using both mechanistic, biophysically motivated computational brain models and data-driven encoders, and (iii) evaluates these latents with unsupervised criteria and supervised downstream tasks. By holding data handling and evaluation constant, we obtain a principled comparison of interpretability, robustness, and task utility across modeling families. Figure 3.1 illustrates the high-level structure of our pipeline. For implementation details, we refer the reader to the repository: https://gitlab.doc.ic.ac.uk/lrh24/thesis.

3.1.1 Problem Statement and Goals

Problem. Given clinical EEG, how do mechanistic and data-driven methods differ in the structure, information content, and downstream utility of their latent representations?
Goals.

1. Design a method-agnostic pipeline that standardizes preprocessing, inputs, and splits.
2. Extract latents from (a) computational brain models (e.g., CTM, Jansen-Rit, Wong-Wang, Hopf) and (b) data-driven methods (e.g., catch22, PCA over PSD, EEGNet/PSD autoencoders). We refer the reader to chapter 2 for detailed descriptions of the methods.
3. Quantify task relevance using fixed classifier/regressor heads on clinical endpoints (abnormality, sex, age), with controlled hyperparameter search.
4. Assess latent quality via unsupervised metrics (independence, clusterability, geometry) and information-theoretic analyses.

3.2 Benchmark dataset: TUH-AB

We use the Temple University Hospital Abnormal EEG Corpus (TUH-AB) as the benchmark dataset throughout this thesis. It provides the clinical realism and scale needed to evaluate both data-driven and mechanistic approaches under consistent conditions. TUH-AB is a curated subset of the larger TUH EEG Corpus, designed for binary screening (normal vs. abnormal), and was introduced by López et al. [18, 63]. In this study, we use version **v3.0.1** (2024-02-07), which standardises header formatting while preserving the original EEG signals.

The dataset is partitioned into training and evaluation splits, each further divided by class label (normal or abnormal). All recordings use an averaged-reference montage, and each EEG file contains a single recording which lasts at least 15 minutes, selected to preserve relevant clinical content. Importantly, each file is labeled at the *session* level, meaning the assigned label reflects the overall clinical impression from that session, rather than moment-to-moment annotations.

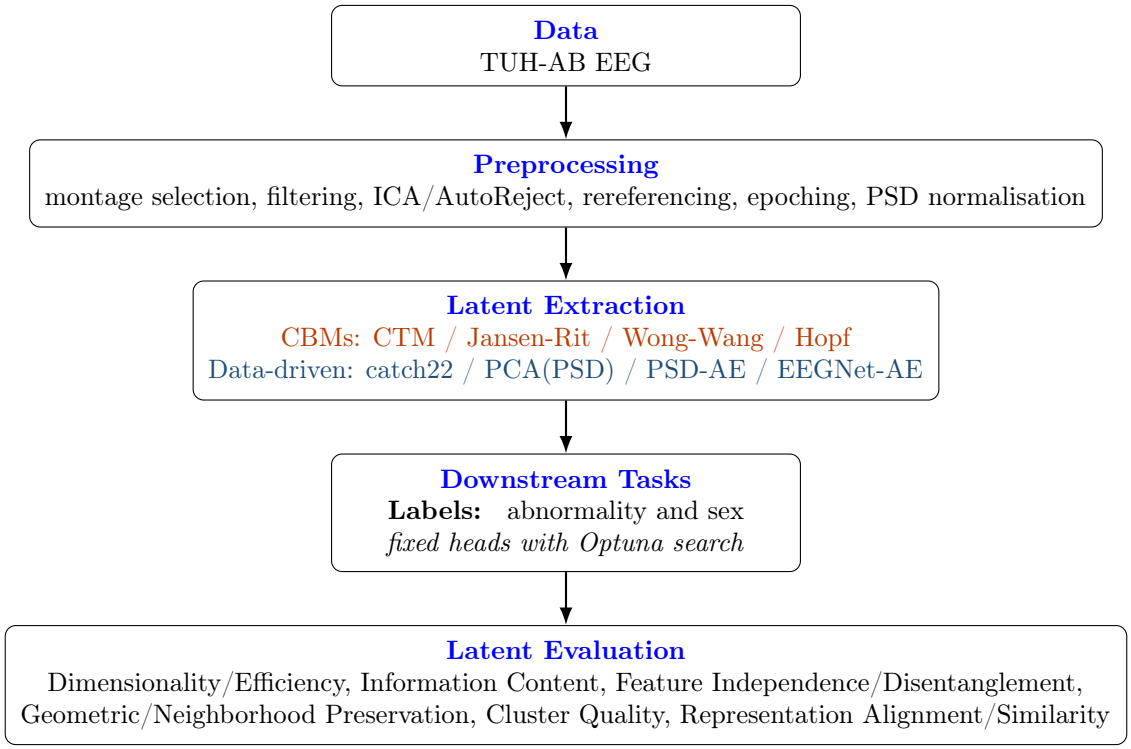


Figure 3.1: High-level pipeline structure: standardised preprocessing, method-agnostic latent extraction, supervised downstream assessment and unsupervised latent evaluation.

Version 3.0.1 includes a total of **2,993** EEG sessions (**1,142 hours**) split into **2,717** training and **276** evaluation examples. These correspond to 2,130 subjects in the training set and 253 in the evaluation set, with subject-wise disjointness enforced between splits. Session labels are roughly balanced within each partition (e.g., 1,371 normal and 1,346 abnormal in the training set). A small number of subjects in the training set have both normal and abnormal sessions, but no subject appears in both train and eval [18, 63].

Manual review in an earlier release (v3.0.0) reported *100%* positive agreement on the evaluation set and *99%* on the training set, indicating strong alignment between corpus labels and visible clinical abnormalities.

The dataset also provides demographic metadata. Subject sex is available and is approximately balanced across splits. For example, in the training set 53% of subjects are female and 47% are male, while in the evaluation set the distribution is 54% female and 46% male. We therefore include **sex** alongside the canonical **abnormality** label in our analyses. By contrast, age information is provided only in aggregate distributions rather than at the level of individual sessions or subjects (see TUH-AB v3.0.1 `AAREADME.txt`). Therefore, we can't use ages as an additional label.

TUH-AB offers key advantages: scale, clinical realism, a unified montage format, and reproducible partitions [18, 63]. At the same time, several caveats apply: labels are only available at the session level, abnormal activity may be temporally sparse within long recordings, and age metadata are provided only in aggregate distributions. While the normal/abnormal and sex distributions are fairly balanced, transparent reporting remains important when working with clinical data. These aspects motivate the use of standardised preprocessing and method-agnostic evaluation pipelines, detailed in Chapter 3. Such preprocessing reduces nuisance variability and highlights clinically relevant spectral structure, ensuring that both data-driven methods and computational brain models can be compared fairly while allowing the latter to focus on the physiological mechanisms they are designed to capture.

3.3 Preprocessing and Standardisation

Robust latent representation learning from clinical EEG hinges on a carefully standardised signal preparation pipeline. Heterogeneity in montages, hardware, and acquisition protocols introduces substantial nuisance variability that can dominate downstream models if not controlled. We

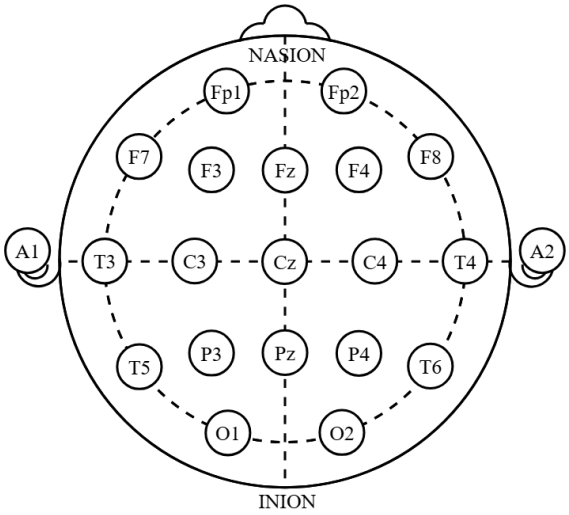


Figure 3.2: Standard 10–20 electrode placement system for EEG[3].

adopt a unified preprocessing stack, implemented primarily with the `MNE-Python` framework [64], that enforces a common channel space and sampling rate, attenuates artifacts while preserving neurophysiological structure, and produces fixed-length epochs suitable for fair, method-agnostic comparison across all extractors.

Channel canonicalisation and montage handling

Clinical EEGs are recorded with electrodes placed according to a specific spatial configuration, known as a *montage*. A montage specifies which electrodes are present and how they are named and arranged on the scalp. In clinical reporting, the term is sometimes also used to denote the reference scheme applied for display, because clinical datasets often vary in naming conventions, auxiliary sensors, and referencing, harmonisation is required. We first align channel labels to the international 10–20 system [65](see Figure 3.2). We then select a fixed canonical subset, in our case we omit A1 and A2, as they are primarily used for referencing in clinical practice rather than for capturing cortical activity, resulting in a 19-channel 10–20 setup. We then impose a deterministic ordering to yield a consistent sensor vector space. Recordings missing any canonical channel are excluded (no interpolation), and a standard 10–20 montage is assigned for spatial consistency. This canonicalisation removes montage-induced confounds and ensures that per-channel methods receive consistent inputs.

Filtering, rereferencing, resampling

To reduce slow drifts and high-frequency noise while preserving conventional EEG rhythm bands, we apply a high-pass filter (1 Hz) and a low-pass filter (45 Hz). This band limitation reflects both practical considerations, higher frequencies are typically dominated by muscle and measurement noise, and theoretical ones, as most computational brain models are formulated to reproduce dynamics in the delta–beta range rather than broadband activity. Power-line contamination, in the US this is at 60 Hz and its harmonic frequencies, is also mitigated by the band-pass filter. Signals are re-referenced to a common average reference to suppress global offsets and improve topographic interpretability. Finally, all recordings are resampled to a common sampling rate f_s ($f_s = 128$ Hz), which standardizes temporal resolution, reduces computational cost, and ensures a consistent frequency resolution, while still avoiding aliasing as the Nyquist–Shannon sampling theorem implies.

Artifact handling (ICA/EOG/ECG, AutoReject, bad-channel interpolation)

Ocular and cardiac artifacts are addressed using an ICA-based procedure[66, 64]: we fit an independent component decomposition on band-limited data with robust amplitude clipping to stabilize the unmixing, identify artifact components via correlations with EOG/ECG channels[64] and

stereotyped time-frequency signatures, and remove them from the data. Muscle bursts and other transient artifacts are annotated using automated detectors (e.g., muscle-zscore heuristics) to mark segments for exclusion. Bad EEG channels are detected via robust dispersion criteria and spatial inconsistency, and we discard the session if that is the case. However, in practice this was never the case as TUH manually reviewed the recordings. At the epoch level, we employ automated rejection using `AutoReject` [67], which learns channel-specific amplitude thresholds by cross-validation and excludes affected epochs. This avoids manual cutoffs and adapts rejection to the statistical properties of each dataset. This layered strategy targets distinct artifact classes while minimising distortion of intact neural activity.

Epoching protocol (length, overlap, masks)

Preprocessed continuous recordings are segmented into fixed-length, non-overlapping windows of duration L seconds (e.g., $L = 10$ s). This window length is commonly used in EEG preprocessing, as it provides sufficient frequency resolution while preserving approximate stationarity of the signal. Fixed-length segmentation ensures comparable input statistics across subjects and facilitates batching on constrained hardware. For each recording, we construct binary masks that exclude windows intersecting annotated bad artifact segments or containing a fraction of missing data. The resulting set of eligible epochs defines the training and evaluation inputs for *all* latent extractors. By fixing epoch length, overlap (here, none), inclusion masks, and split membership a priori, we guarantee that cross-method comparisons are not confounded by differences in data selection. We cap to at most 20 epochs per recording, preferring those with lower beta/alpha power ratio. This balances dataset size and quality control.

PSD estimation

When spectral features are required (either as direct inputs or as intermediate statistics), we estimate power spectral densities (PSDs) using Welch's method[31] with a globally fixed configuration. Spectra are computed on each 10 s epoch with Hann windows of 4 s (512 samples at 128 Hz), 50% overlap, FFT size $N_{\mathrm{FFT}} = 512$, and frequency range restricted to 1–45 Hz. This yields an identical frequency grid across all analyses. For per-channel methods, PSDs are computed channel-wise and concatenated in canonical order. For channel-averaged variants, spectra are averaged across the canonical set. Normalisation is applied in a model-specific manner: mechanistic models operate on unnormalised PSDs and apply log-and-z normalisation internally during fitting, whereas learned baselines use log-and-z normalisation at PSD computation to match their training distribution. These choices produce well-conditioned inputs whose variance is concentrated in physiologically relevant bands while preserving relative spectral shape.

Unless otherwise noted, spectra are compared on the shared Welch frequency grid

$$\mathcal{F} = \{f_1, \ldots, f_K\}, \quad f_1 = 1\,\mathrm{Hz},\ f_K = 45\,\mathrm{Hz},$$

after *log–z normalisation*. For a spectrum $S(f)$ we write

$$\mathrm{norm}\big(S(f)\big) = \frac{\log S(f) - \mu}{\sigma}, \quad \mu = \tfrac{1}{|\mathcal{F}|}\sum_{f\in\mathcal{F}} \log S(f), \quad \sigma^2 = \tfrac{1}{|\mathcal{F}|}\sum_{f\in\mathcal{F}} \big(\log S(f) - \mu\big)^2.$$

Here, $S(f)$ denotes the power spectral density at frequency f, $\mathcal{F}$ is the discrete set of Welch frequency bins between 1 and 45 Hz, μ is the mean log-power across those bins, and σ is the corresponding standard deviation. We use $\mathrm{norm}(\cdot)$ as shorthand throughout this entire chapter.

Choice of aggregated input. For part of our evaluation we require a single PSD representation per subject. There are two common ways to construct such an input: either by selecting the PSD of a specific channel (e.g., O1 [68]) or by averaging across all channels. While individual channels can sometimes capture task-relevant features, they also introduce task- or montage-specific bias. We therefore use the mean PSD across all canonical EEG channels, which avoids privileging any one channel and provides a fair, unbiased representation of the spectral content. This aggregated PSD serves as the basis for methods that operate in lower-dimensional latent spaces.

After applying all preprocessing, artifact rejection, and epoching steps, the pipeline yielded 53622 training and 5459 evaluation standardised 10 s EEG segments across 2717 training and 276 evaluation sessions, which served as the common input set for all latent extractors evaluated in this thesis.

3.4 Latent extraction methods

All extraction methods operate on identically preprocessed inputs (canonical channels, standardised sampling, fixed-length epochs), and produce 1D latent vectors with consistent serialisation for downstream analysis. Subsequent sections detail the extractor architectures, fitting criteria, and hyperparameters, including the amortised CBM variant and followed by a common suite of unsupervised latent diagnostics and supervised head evaluations that enable like-for-like comparisons across families and quantify the empirical gains from amortisation.

3.4.1 Cortico-Thalamic Model (CTM)

Input and preprocessing. We fit CTM parameters to sensor-space power spectral densities (PSDs) computed from identically preprocessed EEG (canonical 10-20 channels, rereferenced, artifact-attenuated, resampled). PSDs are estimated with Welch's method and log-z normalized. per recording. Fitting frequencies are restricted to 1-45 Hz and aligned exactly to the Welch grid to avoid interpolation.

Model and parameters. CTM coarse-grains cortical and thalamic populations into a coupled linear (or weakly nonlinear) neural field whose frequency response $H_{\mathrm{CTM}}(f;\boldsymbol{\theta})$ maps stochastic population drives η to mesoscopic field potentials. We use the 8-parameter Robinson CTM with a fixed spatial grid. The fitted vector is

$$\boldsymbol{\theta} = \big[G_{\mathrm{ee}},\, G_{\mathrm{ei}},\, G_{\mathrm{ese}},\, G_{\mathrm{esre}},\, G_{\mathrm{srs}},\, \alpha,\, \beta,\, t_0\big].$$

The parameters have direct physiological meaning: G_{ee} and G_{ei} set cortical excitation and inhibition gain. G_{ese} and G_{esre} describe corticothalamic loop gain; G_{srs} captures intrathalamic loop gain. α and β are the decay and rise rates of cell-body potentials and t_0 is the corticothalamic loop delay. Together they determine the feedback balance and resonance structure of the model. [2]

Given $\boldsymbol{\theta}$, we compute the analytic CTM PSD on the exact Welch frequency grid shared with the empirical spectra. Absolute scale and offsets are removed by the downstream log–z normalisation used in the loss (so no separate noise amplitude or baseline term is fitted). We consider two variants: (i) channel-averaged PSDs (`ctm_cma_avg`); (ii) per-channel PSDs (`ctm_cma_pc`).

Objective and constraints. Parameters are estimated by minimising mean squared error between normalised empirical and model spectra on the shared Welch grid:

$$\mathcal{L}(\boldsymbol{\theta}) = \frac{1}{|\mathcal{F}|} \sum_{f\in\mathcal{F}} \Big[\mathrm{norm}\big(S_{\mathrm{emp}}(f)\big) - \mathrm{norm}\big(S_{\mathrm{CTM}}(f;\boldsymbol{\theta})\big)\Big]^2,$$

with physiologically motivated box bounds on $\boldsymbol{\theta}$ (including signed ranges where appropriate) and uniform frequency weights.

Estimation. Optimisation uses CMA-ES with bounded domains and tuned convergence criteria (e.g., tolfun and max iterations).

Representation. For `ctm_cma_avg`, the fitted $\boldsymbol{\theta} \in \mathbb{R}^p$ forms the latent vector. For `ctm_cma_pc`, per-channel fits $\boldsymbol{\theta}^{(c)}$ are concatenated in canonical order.

3.4.2 Cortico-Thalamic Model (CTM), amortised inference

Inputs and featurisation. This is a new hybrid approach we designed with inspiration from Momi et al.[51]. We infer CTM parameters from sensor-space power spectral densities (PSDs) computed on identically preprocessed EEG (canonical 10–20 channels, rereferenced, resampled, artifact-attenuated upstream). PSDs are estimated with Welch's method, then transformed by the norm$(\cdot)$ operator (log–z per recording). The CTM forward model is evaluated on the exact Welch frequency grid (derived from the same settings) to avoid interpolation.

Parameterisation. The latent vector is the 8-dimensional CTM parameter set

$$\boldsymbol{\theta} = \big[G_{\text{ee}},\, G_{\text{ei}},\, G_{\text{ese}},\, G_{\text{esre}},\, G_{\text{srs}},\, \alpha,\, \beta,\, t_0\big],$$

which encodes loop gains, dendritic rate constants, and the corticothalamic delay. The physiological interpretation of these parameters is identical to that in the CMA-ES formulation of the CTM. [2]

We report two regimes: channel-averaged (`ctm_nn_avg`) and per-channel (`ctm_nn_pc`); in the latter, a separate $\boldsymbol{\theta}^{(c)}$ is inferred per canonical channel and concatenated in fixed order.

Amortised regressor and training data. We train a feedforward network $g_\phi : \mathbb{R}^K \to \mathbb{R}^8$ that maps a normalised PSD vector $x \in \mathbb{R}^K$ to $\hat{\boldsymbol{\theta}} = g_\phi(x)$. Training data are generated synthetically by sampling $\boldsymbol{\theta}$ from physiologically bounded uniform priors and computing the corresponding analytic CTM PSD on the shared grid. The synthetic spectra are transformed by norm$(\cdot)$ identically to empirical PSDs.

Objective. Our model learns by minimising the mean squared error between the normalised CTM spectrum produced by predicted parameters and the pre-normalised input spectrum x on the same Welch grid:

$$\mathcal{L}(\phi) = \text{MSE}\Big(\text{norm}\big(S_{\text{CTM}}(f;\, g_\phi(x))\big),\; x\Big),$$

restricted to a configured band (1–45 Hz). No parameter-space loss or auxiliary regularisers are used. This gives our implementation an encoder-like structure.

Inference and representation. At test time, we compute the Welch PSD from an empirical epoch, transform it by norm$(\cdot)$, and evaluate $\hat{\boldsymbol{\theta}} = g_\phi(x)$ in a single forward pass (no projection/clamping). For `ctm_nn_avg`, $\hat{\boldsymbol{\theta}}$ is used directly as the latent vector. For `ctm_nn_pc`, per-channel predictions $\{\hat{\boldsymbol{\theta}}^{(c)}\}_c$ are concatenated.

3.4.3 Jansen-Rit Model (JR)

Input and preprocessing. We estimate JR parameters from sensor-space power spectral densities (PSDs) computed on identically preprocessed EEG (canonical 10–20 channels, rereferenced, artifact-attenuated, resampled). PSDs are obtained with Welch's method on a shared 1–45 Hz grid; during fitting, both empirical and model spectra are both log-z normalized.

Model and parameters. The JR model coarse-grains a cortical column into three interacting neural masses (pyramidal, excitatory interneurons, inhibitory interneurons). Linearisation in the frequency domain yields a transfer function $H_{\text{JR}}(f;\boldsymbol{\theta})$ from stochastic synaptic input to pyramidal output. We fit a reduced 6-parameter vector

$$\boldsymbol{\theta} = [C_1,\, A,\, B,\, a,\, b,\, G],$$

where C_1 sets the base pyramidal–excitatory connectivity (with other C_i derived from it), A and B are the maximum amplitudes of excitatory and inhibitory postsynaptic responses, a and b are their characteristic rise/decay rates (time constants), and G is the effective slope of the pyramidal population's linearised input–output function. [69, 26, 11, 23] The fitted spectrum is proportional to $|H_{\text{JR}}(f;\boldsymbol{\theta})|^2$ on the shared frequency grid, with absolute scale removed by normalisation.

Objective and constraints. Parameters are estimated by minimising mean squared error between normalised empirical and model spectra on the shared Welch grid:

$$\mathcal{L}(\boldsymbol{\theta}) = \frac{1}{|\mathcal{F}|} \sum_{f \in \mathcal{F}} \Big[\mathrm{norm}\big(S_{\mathrm{emp}}(f)\big) - \mathrm{norm}\big(S_{\mathrm{JR}}(f;\boldsymbol{\theta})\big) \Big]^2,$$

with physiologically motivated box bounds on $\boldsymbol{\theta}$ and uniform frequency weights.

Estimation. Optimisation uses CMA-ES with bounded domains and tuned convergence criteria (e.g., tolfun and max iterations).

Variants and representation. We consider (i) **channel-averaged** fitting (`jr_avg`), averaging Welch PSDs across canonical channels, and (ii) **per-channel** fitting (`jr_pc`), estimating parameters independently for each channel.

3.4.4 Wong-Wang Model (DMF)

Input and preprocessing. We fit Wong–Wang (dynamic mean-field; DMF) parameters to sensor-space power spectral densities (PSDs) computed from identically preprocessed EEG (canonical 10–20 channels, rereferenced, artifact-attenuated, resampled). PSDs are estimated with Welch's method on a fixed 1–45 Hz grid; before comparison in the loss, spectra are log–z normalised per vector to stabilise scale.

Model and parameters. We use a single-node dynamic mean-field (DMF) formulation of the Wong–Wang model simulated with Euler–Maruyama.[70, 13] The state is the NMDA gating variable $S(t) \in [0, 1]$ with dynamics

$$\dot{S}(t) = -\frac{S(t)}{\tau_s} + \big(1 - S(t)\big)\, \gamma\, H\big(x(t)\big) + \sigma\, \xi(t), \quad x(t) = J\, S(t) + I_0,$$

where $\xi(t)$ is unit-variance Gaussian white noise and $H(x)$ is the standard DMF nonlinearity in (a, b, d) form (constants fixed).

The fitted parameter vector is

$$\boldsymbol{\theta} = \big[J,\, \tau_s,\, \gamma,\, I_0,\, \sigma\big].$$

Given $\boldsymbol{\theta}$, we simulate $S(t)$ and compare its Welch PSD to the empirical PSD on the shared frequency grid.

Objective and constraints. Parameters are estimated by minimising mean squared error between normalised empirical and model spectra on the shared Welch grid:

$$\mathcal{L}(\boldsymbol{\theta}) = \frac{1}{|\mathcal{F}|} \sum_{f \in \mathcal{F}} \Big[\mathrm{norm}\big(S_{\mathrm{emp}}(f)\big) - \mathrm{norm}\big(S_{\mathrm{DMF}}(f;\boldsymbol{\theta})\big) \Big]^2,$$

with physiologically motivated box bounds on $\boldsymbol{\theta}$ and uniform frequency weights.

Estimation. Optimisation uses CMA-ES with bounded domains and tuned convergence criteria (e.g., tolfun and max iterations).

Variants and representation. We consider (i) **channel-averaged** fitting (`wong_wang_avg`), comparing to the channel-averaged PSD, and (ii) **per-channel** fitting (`wong_wang_pc`), estimating parameters independently per channel.

3.4.5 Hopf (Stuart–Landau) oscillator

Input and preprocessing. We estimate Hopf-based spectral parameters from sensor-space power spectral densities (PSDs) computed on identically preprocessed EEG (canonical 10–20 channels, rereferenced, artifact-attenuated, resampled). PSDs are estimated with Welch's method on a fixed 1–45 Hz grid; before comparison in the loss, spectra are log–z normalised per vector to stabilise scale.

Model and parameters. As a normal-form approximation to oscillatory dynamics near a supercritical Hopf bifurcation, a Lorentzian provides a good local model of narrowband peaks. We fit, *independently within each canonical band* $b \in \mathcal{B} = \{\delta(1-4),\ \theta(4-8),\ \alpha(8-13),\ \beta(13-30)\}$, a Lorentzian-with-baseline

$$L_b(f) \;=\; \frac{A_b}{(f - f_{0,b})^2 + \gamma_b^2} \;+\; b_b,$$

with parameters $\boldsymbol{\theta}_b = \{A_b, f_{0,b}, \gamma_b, b_b\}$ where $A_b \geq 0,\ f_{0,b} \in [f_{\min,b}, f_{\max,b}],\ \gamma_b > 0$, and $b_b \geq 0$.[14] The final feature vector concatenates $\boldsymbol{\theta}_b$ over bands in a fixed order.

Objective and constraints. Within each band window we minimize a least-squares discrepancy on the *normalized* PSD,

$$\min_{A_b,\, b_b} \left\| y(f) - \big(A_b\, k_{f_{0,b},\gamma_b}(f) + b_b\big) \right\|_2^2,$$

subject to $A_b \geq 0,\ b_b \geq 0,\ f_{0,b} \in [f_{\min,b}, f_{\max,b}]$, and $\gamma_b \in [\gamma_{\min}, \gamma_{\max}]$. Here $y(f)$ is the log–z normalised PSD and $k_{f_{0,b},\gamma_b}(f) = ((f - f_{0,b})^2 + \gamma_b^2)^{-1}$. No additional frequency weighting is used.

Estimation. We perform a deterministic grid search over $(f_{0,b}, \gamma_b)$ within band bounds; for each pair we solve the linear least-squares problem in closed form for (A_b, b_b) and select the minimum-SSE fit. Frequency grid, normalisation, and bounds are identical across subjects for comparability.

Variants and representation. We consider (i) **channel-averaged** fitting (`hopf_avg`), where Welch PSDs are averaged across canonical channels for robustness, and (ii) **per-channel** fitting (`hopf_pc`), estimating parameters independently to retain spatial specificity. The fitted parameters $\{A_b, f_{0,b}, \gamma_b, b_b\}_{b \in \mathcal{B}}$ (concatenated across bands and channels in `hopf_pc`) define 1D mechanistic latent vectors used unchanged by the common latent diagnostics and downstream heads.

3.4.6 catch22

Input and preprocessing. We extract *catch22[1]* features from identically preprocessed, artifact-attenuated, rereferenced, and resampled EEG epochs using the canonical 10–20 channel ordering. Inputs are fixed-length time-domain segments at a standardised sampling rate; no additional per-epoch z-scoring is applied.

Feature mapping. For each canonical channel $c \in \mathcal{C}$ (19-channel set), we compute the 22 canonical time-series characteristics using the reference *catch22* implementation, yielding a per-channel vector $\mathbf{z}^{(c)} \in \mathbb{R}^{22}$ that summarizes distributional shape, linear and nonlinear autocorrelation structure, stationarity/trend, entropy/complexity, motifs/periodicity, and transition/outlier statistics. The subject-level latent is the concatenation

$$\mathbf{z} \;=\; \mathrm{concat}\big(\mathbf{z}^{(c)}\big)_{c \in \mathcal{C}} \in \mathbb{R}^{22 \times |\mathcal{C}|},$$

preserving channel topology through a fixed ordering. To maintain a consistent dimensionality under occasional channel omissions, we insert a zero vector in place of missing channels.

Normalisation and stability. Because *catch22* includes heterogeneous functionals (with distinct units and ranges), we retain raw feature values (no additional feature-wise normalisation here); downstream models may apply per-coordinate standardisation within the training split. The mapping is deterministic and parameter-free, ensuring exact reproducibility given identical inputs.

Variants and representation. We use a single, per-channel variant (no averaging), prioritising spatial specificity and post hoc interpretability. The resulting 1D latent vectors (length 22×19) are serialised without further transformation and forwarded unchanged to the common latent diagnostics (independence, clusterability, geometry) and supervised heads.

3.4.7 PCA over PSD

Input and preprocessing. We derive principal-component latents from sensor-space power spectral densities (PSDs) computed on identically preprocessed EEG (canonical 10-20 channels, rereferenced, artifact-attenuated, resampled). PSDs are estimated with Welch's method on a fixed 1-45 Hz grid and log–z normalised per vector to stabilize scale.

Training and freezing. We fit a StandardScaler + PCA on the *training split only*. Training data consist of per-channel PSD vectors stacked across all canonical channels and recordings, yielding a design matrix $X \in \mathbb{R}^{n \times d}$ (one row per channel instance; d frequency bins). After standardisation $X_s = (X - \mu)/\sigma$, PCA is trained to obtain the top k components $V \in \mathbb{R}^{k \times d}$ (whitening optional; we use non-whitened components). The frozen artifact stores μ, σ, V (and eigenvalues) for deterministic reuse across subjects.

Runtime extraction. Given an input PSD vector $x \in \mathbb{R}^d$, we compute the latent

$$\mathbf{z} \;=\; V\left(\frac{x-\mu}{\sigma}\right) \in \mathbb{R}^k.$$

Two variants are supported: (i) **per-channel** (`pca_pc`), where the transform is applied to each channel PSD and the k-dimensional codes are concatenated in canonical order (latent length $k \times |\mathcal{C}|$); and (ii) **channel-averaged** (`pca_avg`), where channel PSDs are averaged prior to transformation (latent length k).

Dimensionality and parameters. We fix k a priori (e.g., k=8) to balance variance capture and compactness; k is chosen on the training split (explained-variance curves reported). No label information is used during fitting. Frequency grid, normalisation, and component matrices are identical across subjects to ensure comparability.

Representation and reproducibility. Latents are 1D float vectors serialised without further processing and passed unchanged to the common latent diagnostics and downstream heads. Using a frozen scaler and projection guarantees exact reproducibility and prevents train–test contamination, enabling like-for-like comparison with mechanistic and learned extractors under identical inputs and splits.

3.4.8 PSD Autoencoder (PSD-AE)

Input and preprocessing. We learn latents from sensor-space power spectral densities (PSDs) computed on identically preprocessed EEG (canonical 10–20 channels, rereferenced, artifact-attenuated, resampled). PSDs use a shared Welch configuration (1–45 Hz band, Hann, $N_{\text{per\_seg}}$=512, 50% overlap, N_{FFT}=512). Each PSD vector is log-transformed and z-scored per vector; the model operates directly on these normalised PSDs.

Architecture. A lightweight encoder–decoder in PSD space: the encoder maps a d-dimensional PSD to a k-dimensional bottleneck $\mathbf{z} \in \mathbb{R}^k$ via stacked affine layers with nonlinearities; the decoder mirrors this mapping to reconstruct the input PSD.

Objective and training protocol. Training uses the *training split only*. From time-domain segments we compute per-channel Welch PSDs and apply per-vector log–z normalisation. The autoencoder minimizes mean-squared reconstruction error (MSE) on these normalised PSDs:

$$\mathcal{L}_{\text{AE}} \;=\; \|\hat{\mathbf{x}} - \mathbf{x}\|_2^2.$$

Optimisation uses Adam with early stopping on a fixed random validation subset (seeded). Frequency grid, normalisation, batch size, and seeds are held constant. After convergence, weights are frozen for feature extraction.

Variants and runtime extraction. (i) **Channel-averaged** (`psd_ae_avg`): average channel PSDs before encoding to yield a k-dimensional latent. (ii) **Per-channel** (`psd_ae_pc`): encode each channel PSD and concatenate the k-dimensional codes in canonical channel order, producing a latent of length $k \times |\mathcal{C}|$ (with $|\mathcal{C}|$=19). In both cases, the latent is the encoder bottleneck $\mathbf{z}$; no task supervision or fine-tuning is applied.

Representation and comparability. Latents are serialised as 1D float vectors and passed unchanged to the common latent diagnostics and supervised heads. A frozen encoder, shared Welch grid, and identical normalisation ensure like-for-like comparison across methods. The bottleneck size k is fixed a priori (reported in the setup) to balance compactness and fidelity.

3.4.9 EEGNet Autoencoder

Input and preprocessing. We learn latents from multichannel time-domain EEG epochs that share the common preprocessing stack (canonical 10–20 channels, rereferenced, artifact-attenuated, resampled). Let $\mathbf{x} \in \mathbb{R}^{C \times T}$ denote a standardised segment (channels C, samples T).

Architecture. The encoder follows an EEGNet-style design with depthwise-separable convolutions: an initial temporal convolution (band-pass–like filtering), a depthwise spatial convolution (channel-wise topographies), and a separable convolution block that increases capacity at low parameter cost. The encoder maps $\mathbf{x} \mapsto \mathbf{z} \in \mathbb{R}^k$. The decoder mirrors this structure via upsampling and convolutions to reconstruct $\hat{\mathbf{x}}$.

Objective. Training minimizes a spectral reconstruction loss on Welch power spectra:

$$\mathcal{L} \;=\; \left\| \tilde{P}(\hat{\mathbf{x}}) - \tilde{P}(\mathbf{x}) \right\|_2^2,$$

where $P(\cdot)$ denotes Welch PSDs computed on the shared frequency grid (1–45 Hz; fixed parameters across subjects), and $\tilde{P}$ denotes per-vector normalisation (unit-sum scaling across frequency bins followed by per-vector standardisation). Unlike norm$(\cdot)$, which applies log–z scaling, $\tilde{P}$ was chosen here to stabilize training in the convolutional autoencoder. No additional time-domain or time–frequency loss terms are used. Weight decay is applied via the optimizer.

Training protocol. Optimisation uses Adam on the *training split only* with a fixed batch size, a held-out validation subset for early stopping, and deterministic seeds. Input shapes and Welch parameters are held constant for comparability. After convergence, the encoder is frozen and used solely for feature extraction.

Runtime extraction and representation. At inference, the latent vector is the encoder bottleneck $\mathbf{z} \in \mathbb{R}^k$ (no label supervision or fine-tuning). We operate on full multichannel inputs (no channel averaging); the latent length k is fixed a priori and reported alongside reconstruction metrics.

3.4.10 Fit quality of CBM inversions

To confirm that inversion produced meaningful embeddings, we evaluated the spectral reconstruction quality of fitted CBMs. CMA-ES was configured with model-specific parameter bounds, a maximum of 600 iterations, and a function tolerance of 10^{-4}. This choice ensured that we utilize the potential of each method, while keeping the optimisation computationally feasible for large-scale evaluation. Typical runs converged within a few hundred iterations and captured the dominant alpha and beta components, although residual mismatches were common at higher frequencies. Figure 3.3 illustrates a representative cortico–thalamic fit obtained with CMA-ES, while the Appendix A shows sample fits for every CBM.

The amortised CTM regressor dispenses with iterative optimisation, producing parameter estimates in a single forward pass. Across splits it yielded stable reconstructions with consistently low error, and captured the main spectral peaks of the empirical PSDs. An example is shown in Fig. 3.4, where the regressor approximates the spectral structure with accuracy comparable to CMA-ES.

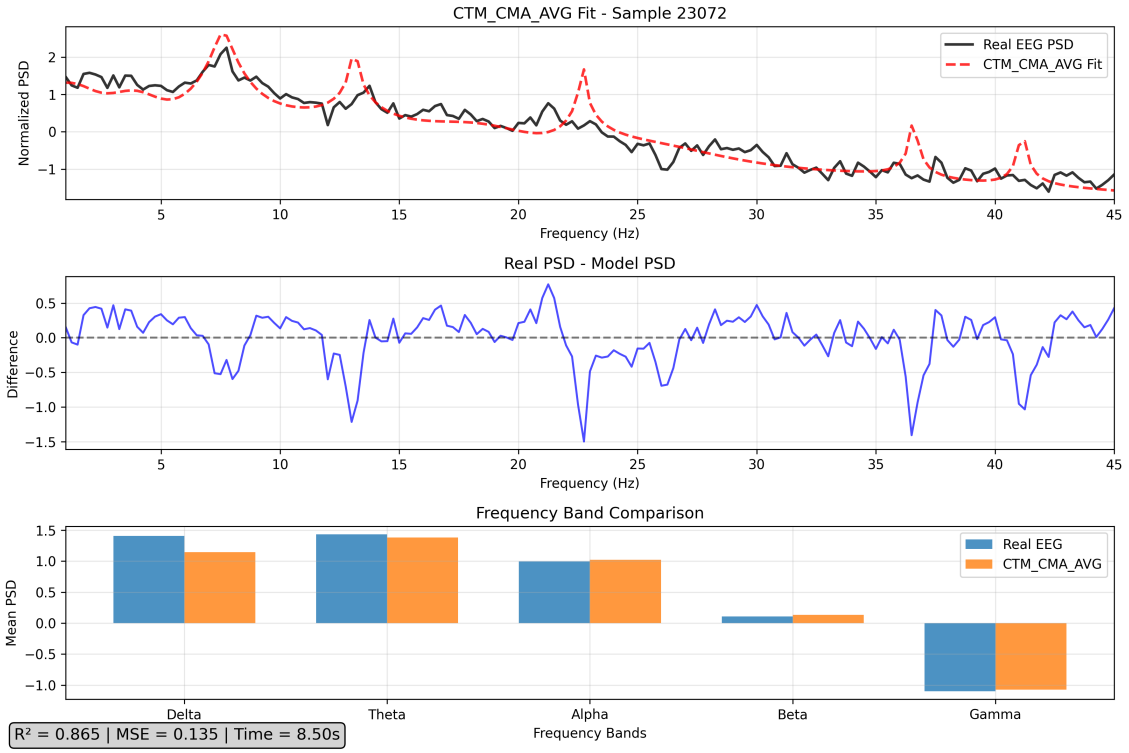


Figure 3.3: Comprehensive ctm_cma_avg fitting result for sample 23072. The plot illustrates the empirical PSD (black) against the fitted model spectrum (red), with residuals and frequency-band decompositions shown in the lower panels.

Jansen–Rit and Hopf models similarly captured narrowband alpha peaks, while Wong–Wang fits tended to produce broader spectra. Overall, these fittings provided a reliable and computationally efficient basis for subsequent latent-space analyses, even if not every spectral detail was recovered.

3.5 Downstream tasks

To assess how much task-related information is preserved in the learned representations, we train lightweight prediction heads on top of fixed latents and evaluate them on held-out data. Heads are fit only on training latents, with early stopping guided by a validation split, and then assessed once on the evaluation latents. All procedures are kept identical across methods to allow direct comparisons.

Tasks and labels

Two binary endpoints from the TUH dataset are considered:

- **Abnormal EEG** distinguishing normal from abnormal recordings
- **Sex** encoded as $\{0, 1\}$ with $0 =$ male and $1 =$ female

Age is available but omitted from supervised evaluation here. The pipeline supports both classification and regression tasks, though only classification is used in this work. Labels are carried alongside the latents during extraction but are never used by the representation models themselves.

Prediction heads and feature scaling

For each task, an independent feedforward network is trained with a ReLU trunk and a single scalar output:

$$\text{MLP}: \ \mathbb{R}^d \xrightarrow{(\text{Linear}+\text{BN}+\text{ReLU}+\text{Dropout})\times L} \mathbb{R}.$$

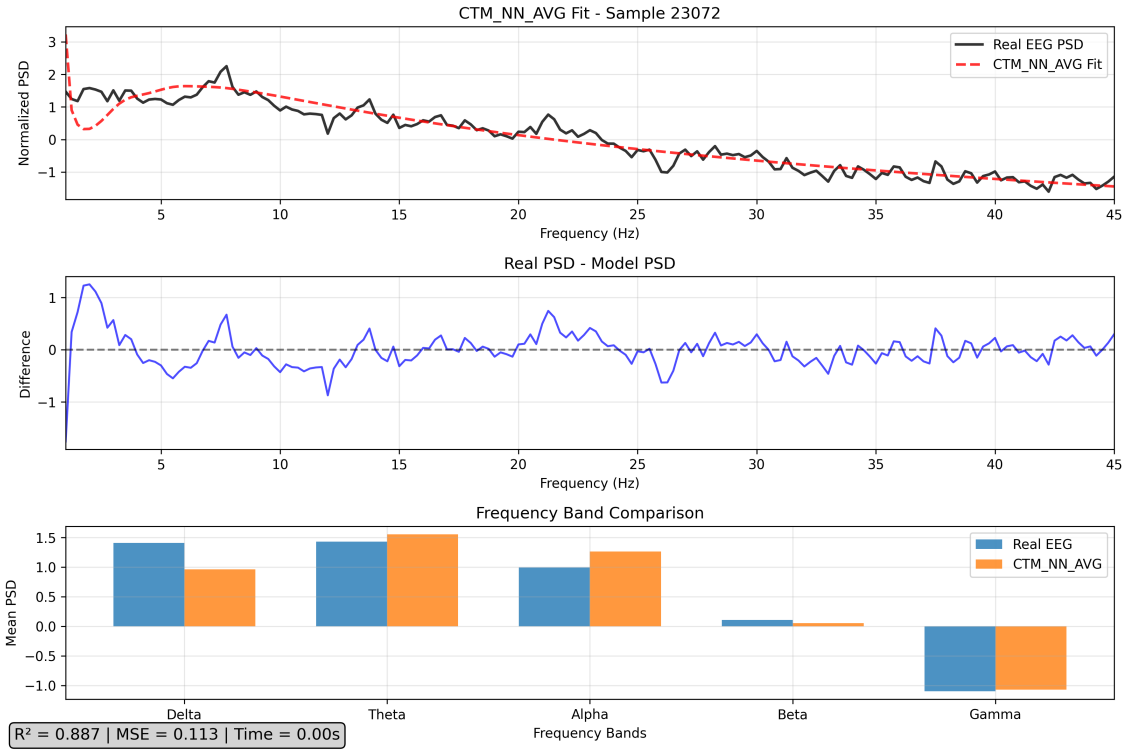


Figure 3.4: Example fit from the amortised CTM regressor (sample 23072). The model captures the dominant spectral peaks and overall PSD structure in a single forward pass, achieving accuracy comparable to CMA-ES.

This design applies the same building block L times, reducing width across layers and producing one output per task. Classification is carried out using raw logits and optimised with BCEWithLogitsLoss, applying a fixed threshold of 0.5 during inference. No class weighting is introduced. Before training, latent features are standardised using statistics computed on the training split, and the same transformation is consistently applied to validation and evaluation latents.

Hyperparameter search

Hyperparameters are tuned with a seeded TPE-based search, performed separately for each method and task. We minimize the validation loss (Binary Cross Entropy). The search space includes:

- Depth $L \in \{2, 3, 4\}$
- Base width $b \in \{64, 128, \ldots, 512\}$ with per-layer halving and a minimum of 16 units
- Learning rate in $[10^{-5}, 10^{-3}]$ (log-uniform), dropout in $[0, 0.3]$, weight decay in $[10^{-6}, 10^{-2}]$ (log-uniform)
- Scheduler choice between Reduce-on-Plateau, Cosine with warm restarts, or None

Each trial is trained for up to 100 epochs with early stopping, requiring at least five epochs before selection. All searches and training runs are seeded for reproducibility.

Training, evaluation, and reproducibility

The training latents are divided into a subject-wise training and validation split using `GroupShuffleSplit`. This ensures that no subject contributes samples to both splits, thereby preventing leakage and stabilising model selection. The same split is reused across all trials within a given method–task pair. Within each trial, the model with the lowest validation loss is retained, and across all trials the overall best configuration is selected for final evaluation.

Table 3.1: Optuna and training settings used throughout the downstream evaluations.

Setting	Value
Trials per task	50
Validation fraction	0.15
Training split size (epochs)	~ 45579
Validation split size (epochs)	~ 8043
Evaluation size (epochs)	5459
Early stopping patience	7
Batch size	512
Optuna seed	42
Min. epochs before selection	5
Schedulers	plateau, cosine, none
Max epochs	100
Learning rate range	10^{-5} to 10^{-3}
Dropout range	0 to 0.3
Weight decay range	10^{-6} to 10^{-2}

After selecting the best hyperparameters, we refit the readout on the training latents and evaluate on the held-out split. Standardisation parameters are estimated on the training split and applied unchanged to validation and test data.

3.6 Latent Evaluation

We evaluate latent representations $Z \in \mathbb{R}^d$ produced by all extractors under a single, method-agnostic protocol. Unless stated otherwise, computations are performed on the evaluation split with identical inputs, frequency grids, and seeds. Latents are cached per method alongside sample IDs.

For each method and split we construct

$$\mathcal{Z}_{\text{split}} = \{(\mathbf{z}_i,\, y_i^{\text{sex}},\, y_i^{\text{abn}},\, \texttt{sample\_id}_i)\}_{i=1}^{N_{\text{split}}}.$$

Latents $\mathbf{z}_i$ are 1D float vectors saved with their labels and sample IDs.

Dimensionality and Efficiency

Variance and active units. For each split (train/eval), we treat latents as rows of a matrix $Z \in \mathbb{R}^{N \times d}$ and compute per-dimension variance with population normalisation:

$$v_j \;=\; \mathrm{Var}[Z_j] \quad \text{(computed with } \texttt{unbiased=False}\text{)}.$$

We define *active units* as dimensions with variance above a fixed threshold, i.e.

$$\#\{\text{active}\} \;=\; \#\{\, j \;:\; v_j > 10^{-3} \,\}.$$

We export variance histograms per split and a train–eval variance comparison scatter to visualize dimension utilisation and generalisation.

Explained variance and effective dimensionality. On train latents only, we fit PCA (components up to $\min(N_{\text{train}}, d)$) and report the explained-variance ratio vector $\big(\mathrm{EVR}_k\big)_{k\geq 1}$ together with the sum of the top-5 ratios $\sum_{k=1}^{5} \mathrm{EVR}_k$. We also plot effective dimensionality curves by counting raw latent dimensions above a variance threshold across a logarithmic grid $\tau \in [10^{-6}, 10^{-1}]$,

$$\mathrm{ED}(\tau) \;=\; \#\{\, j \;:\; v_j > \tau \,\},$$

for both train and eval splits, summarising capacity usage beyond linear PCA variance.

Information Content

For each available target $Y \in \{\text{sex}, \text{abnormal}\}$, we estimate per-dimension mutual information $I(Z_j; Y)$ using scikit-learn's `mutual_information_classif`. When labels are encoded as $\{1, 2\}$, they are remapped to $\{0, 1\}$ via $\mathbb{1}[y{=}2]$. If fewer than two classes are present in a split, the metric is skipped. We report the mean mutual information across latent dimensions and also retain the full per-dimension MI vector for each split. All computations are performed independently per method under the same splits and inputs.

Feature Independence / Disentanglement

Pairwise HSIC with median-heuristic bandwidths. We standardize each latent coordinate to zero mean and unit variance and compute a biased Hilbert–Schmidt Independence Criterion (HSIC) between all pairs of dimensions using Gaussian kernels. For a single coordinate Z_j, we form the kernel

$$K_{ab}^{(j)} \;=\; \exp\Big(-\tfrac{(z_{aj}-z_{bj})^2}{2\sigma_j^2}\Big), \quad \sigma_j \;=\; \sqrt{\tfrac{1}{2}\,\text{median}\{(z_{aj}-z_{bj})^2 : a \neq b\} + 10^{-7}}.$$

Each kernel is centered via $K_c^{(j)} = HK^{(j)}H$ with $H = I - \frac{1}{n}\mathbf{1}\mathbf{1}^\top$. For dimensions $i \neq j$, we compute

$$\text{HSIC}(i,j) \;=\; \frac{\langle K_c^{(i)},\, K_c^{(j)}\rangle_F}{(n-1)^2},$$

yielding a symmetric $d \times d$ matrix with zeros on the diagonal; coordinates that are (numerically) constant (standard deviation $< 10^{-6}$) are assigned zero HSIC. We report the full HSIC matrix and a global disentanglement score defined as the mean off-diagonal HSIC (lower is better). For efficiency, we subsample up to 10,000 samples per split with a fixed seed (42) when n is large, and save per-split heatmaps.

Geometric / Neighborhood Preservation

We evaluate geometric preservation by comparing each method's latent space Z to a *power–spectral–density* (PSD) representation of the same epochs. In this context, we treat PSDs as the *original space*: it provides the reference geometry against which all embeddings are assessed. PSD is a stable epoch-level summary of oscillatory content, aligns with how most spectral models are constructed, and is interpretable along frequency and channel axes. Using PSD as the reference also makes comparisons across heterogeneous methods fair, even when their training inputs differ, and it allows us to work with correlation distances that are robust to global power scaling and referencing differences.

Reference space. On the held-out evaluation split, we compute Welch PSDs (1–45 Hz band, $n_{\text{fft}}{=}512$, $n_{\text{per\_seg}}{=}512$, $n_{\text{overlap}}{=}256$) for the canonical 19 EEG channels. After base-10 log scaling and per-channel z-scoring, two versions are derived depending on model group:

- **Small group:** channel spectra are averaged to form a single vector per epoch (aggregated PSD).
- **Medium group:** channel spectra are stacked to preserve per-channel structure.

Flattened PSD features yield $X_{\text{PSD}} \in \mathbb{R}^{N \times p}$. Latents $Z \in \mathbb{R}^{N \times d}$ and PSD features are aligned strictly by `sample_id`, ensuring one-to-one comparison of the same EEG epochs. For efficiency, metrics are computed on up to 5,000 samples with a fixed seed of 42 when N is large.

Metrics. We quantify neighborhood and global distance preservation of Z with respect to the PSD reference using three standard measures:

- **Trustworthiness** ($k{=}10$). Penalizes intrusions of non-neighbors when moving from PSD to latent space:

$$T \;=\; 1 - \frac{2}{nk(2n-3k-1)} \sum_{i=1}^{n} \sum_{j \in \mathcal{N}_k^{(Z)}(i)} \max\{0,\, r_{ij}^{(\text{PSD})} - k\},$$

where $r_{ij}^{(\mathrm{PSD})}$ is the rank of j among i's neighbors in PSD space, and $\mathcal{N}_k^{(Z)}(i)$ is the k-NN set around i in latent space. High trustworthiness (≈ 1) means the latent rarely creates false neighbors.

- **Continuity** (k=10). Penalizes exclusions of true PSD neighbors in latent space:

$$C = 1 - \frac{2}{nk(2n-3k-1)} \sum_{i=1}^{n} \sum_{j \in \mathcal{N}_k^{(\mathrm{PSD})}(i)} \max\{0,\, r_{ij}^{(Z)} - k\},$$

where $r_{ij}^{(Z)}$ is the rank of j in latent space. High continuity means the latent space does not split apart neighborhoods present in the PSD manifold.

- **Distance correlation**. Pearson correlation between the vectorised upper-triangular pairwise distance matrices in PSD versus Z. This captures global structure: values near 1 indicate that inter-sample distances are preserved across the entire dataset, not just locally.

Together, trustworthiness emphasizes avoidance of false neighbors, continuity emphasizes retention of true neighbors, and distance correlation emphasizes preservation of global geometry. Their combination provides a balanced assessment of local and global fidelity.

Cluster Quality

To probe the structure of the latent space without using labels, we evaluate how well it supports unsupervised partitioning. For each split (train and evaluation), we run k-means clustering on the latent matrix $Z \in \mathbb{R}^{N \times d}$ using scikit-learn with k=5, `n_init`=10, and `random_state`=42, applied to all samples without subsampling. The choice $k = 5$ balances resolution and stability: it is large enough to reveal non-trivial subgroup structure, yet small enough to avoid overfragmentation or producing clusters that are too sparse. It also provides a consistent reference granularity across methods and splits, allowing direct comparison of how well latent geometry supports coherent groupings.

If $N \leq k$ or the fitted assignment collapses to a single cluster, metrics are skipped as degenerate. Otherwise, we compute three standard cluster validity indices directly on Z and the k-means labels:

- **Silhouette score**: measures how close each point is to its assigned cluster compared to other clusters. Higher values indicate clusters that are both compact and well separated.
- **Davies–Bouldin index**: computes the ratio of within-cluster scatter to between-cluster separation, averaged across clusters. Lower values indicate more distinct and less overlapping clusters.
- **Calinski–Harabasz index**: compares the dispersion of points within clusters to the dispersion between cluster centroids. Higher values indicate tighter, more well-defined clusters relative to their separation.

Together, these scores capture complementary aspects of latent structure: how compact clusters are, how distinct they are from one another, and how well they balance within-cluster and between-cluster variance. All evaluations are performed independently per method under identical splits and preprocessing.

Representation Alignment / Similarity

For pairwise comparisons between two methods on the *same* split, we first align samples by intersecting `sample_id`s and row matching the latent matrices. Let the aligned embeddings be $Z_1 \in \mathbb{R}^{n \times d_1}$ and $Z_2 \in \mathbb{R}^{n \times d_2}$, where n is the number of common samples. We then compute:

- **Linear CKA** *(global linear similarity between representations).*
  Measures how similar the centered Gram matrices are, hence how much the two representations encode the same linear relationships across samples. Invariant to orthogonal transforms and isotropic scaling of features. Values near 1 indicate strong alignment. With row centering $X_c = X - \bar{X}$, define $K = X_c X_c^\top$, $L = Y_c Y_c^\top$, $H = I - \frac{1}{n}\mathbf{1}\mathbf{1}^\top$, and $K_H = HKH$, $L_H = HLH$. The similarity is

$$\mathrm{CKA}_{\mathrm{lin}}(Z_1, Z_2) = \frac{\langle K_H,\, L_H \rangle_F}{\sqrt{\langle K_H, K_H \rangle_F \, \langle L_H, L_H \rangle_F}} \in [-1, 1].$$

- **RBF CKA** *(nonlinear similarity via kernels).*
  Extends CKA with Gaussian kernels to capture smooth nonlinear correspondences between sample relationships. Bandwidths are set by the median heuristic for each space. Let

$$K_{ab} = \exp\big(-\gamma_1 \|z_a^{(1)} - z_b^{(1)}\|^2\big), \quad L_{ab} = \exp\big(-\gamma_2 \|z_a^{(2)} - z_b^{(2)}\|^2\big),$$

  with $\gamma_i = \big(2\,\text{median}\{\|z_a - z_b\|^2 : a \neq b\}\big)^{-1}$. Apply the same centered alignment as above and compute CKA on K and L.

- **CCA max correlation** *(strongest shared linear mode across subspaces).*
  Canonical Correlation Analysis finds pairs of linear combinations, one in each representation space, that are maximally correlated. The algorithm centers both representations and identifies canonical directions through eigendecomposition of cross-covariance matrices. We report the maximum absolute canonical correlation over $k = \min(d_1, d_2, n-1)$ components. Values are in $[0, 1]$. High values indicate strong shared linear structure between representations.

- **Distance geometry correlation** *(agreement of pairwise distances).*
  Tests whether the two spaces induce similar inter-sample distances, independent of rotations or translations. Let D_1, D_2 be Euclidean distance matrices of Z_1, Z_2. Report the Pearson correlation between the vectorised upper triangles of D_1 and D_2. Values near 1 indicate similar global geometry.

- **k-NN Jaccard overlap** *(local neighborhood agreement).*
  For each sample, take its k nearest neighbors in Z_1 and Z_2 under Euclidean distance, compute the Jaccard index of the two neighbor sets, then average across samples. We use k=10 by default. Scores lie in $[0, 1]$ and emphasize local structure.

- **Procrustes disparity** *(near isometry after optimal alignment).*
  After optimal similarity transform that includes scaling, rotation, and translation, Procrustes reports the residual mismatch. Lower is better and zero means the embeddings are equivalent up to that transform. If $d_1 \neq d_2$, both embeddings are first projected with PCA to the common dimension $k = \min(d_1, d_2)$.

All metrics are aligned by `sample_id`, with identical splits and inputs across methods.

Chapter 4

Results

This chapter presents a systematic comparison of unsupervised latent representations across mechanistic computational brain models (Cortico–Thalamic, Jansen–Rit, Wong–Wang, Hopf) and data-driven baselines (catch22, PCA over PSD, PSD–AE, EEGNet–AE). All methods are evaluated within the unified pipeline described in Chapter 3, with identical preprocessing, Welch frequency grid, and subject-disjoint train/validation/test splits. This ensures that observed differences arise from the representations themselves rather than implementation details.

The goal of this chapter is twofold. First, we report how well the representations support classification of abnormal vs. normal EEG (TUH-AB) and subject sex. Accuracy on held-out data is the primary metric. Second, we provide a detailed analysis of latent space quality, covering dimensionality efficiency, information content, independence, geometry preservation, cluster quality, and representational similarity across methods. Taken together, these perspectives link predictive utility with structural properties of the representations.

To aid comparability we group methods by *weight class*, which reflects input granularity and latent size: **Small** uses 5–16 latent dimensions and operates on an aggregated power spectrum across channels. **Medium** uses 114–418 latent dimensions and operates on per-channel spectra or raw inputs.

Within each weight class, the following models are evaluated:

Small (aggregated channels)

- `ctm_cma_avg`: Cortico–Thalamic Model (CMA–ES fit), 8 latent dimensions.
- `ctm_nn_avg`: Cortico–Thalamic Model with amortised regressor, 8 latent dimensions.
- `hopf_avg`: Hopf oscillator, 16 latent dimensions.
- `jr_avg`: Jansen–Rit model, 6 latent dimensions.
- `wong_wang_avg`: Wong–Wang model, 5 latent dimensions.
- `pca_avg`: PCA over average PSD, 8 latent dimensions.
- `psd_ae_avg`: PSD autoencoder, 8 latent dimensions.

Medium (per-channel or raw inputs)

- `ctm_nn_pc`: Cortico–Thalamic Model with amortised regressor, 114 latent dimensions.
- `hopf_pc`: Hopf oscillator, 152 latent dimensions.
- `jr_pc`: Jansen–Rit model, 114 latent dimensions.
- `pca_pc`: PCA over per-channel PSD, 152 latent dimensions.
- `psd_ae_pc`: PSD autoencoder, 152 latent dimensions.
- `c22`: Catch22 feature vector from per-channel raw data, 418 latent dimensions.
- `eegnet`: EEGNet-based autoencoder from raw data, 128 latent dimensions.

The following sections present results grouped by weight class, beginning with the small, aggregated representations before turning to the medium, per-channel or raw-input methods. For additional latent space evaluation plots and metrics, we refer the reader to the Appendix A.

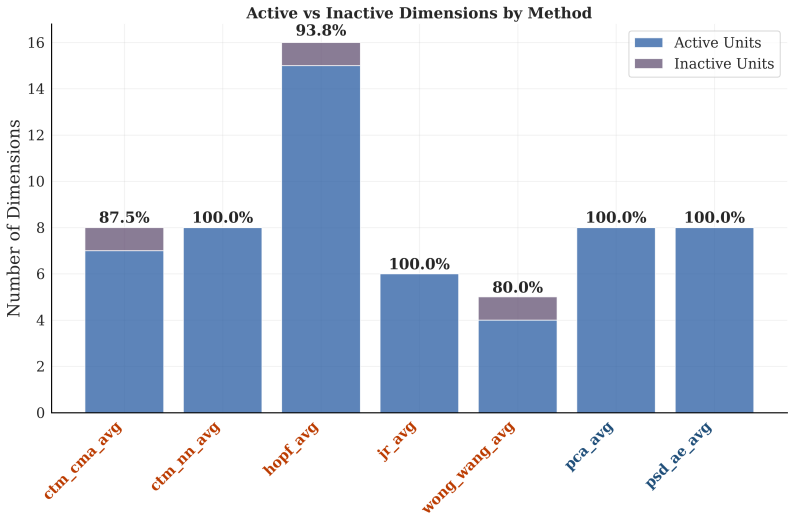


Figure 4.1: Dimensionality efficiency for the Small group methods. Each bar shows the number of active and inactive latent dimensions per method, with dimensions flagged as inactive if they explain less than 10^{-3} variance. This highlights differences in how efficiently each method utilizes its latent capacity.

4.1 Weight Class - Small

For the *Small* weight class, we group and evaluate seven methods that operate on averaged power spectra with compact latent spaces (5–16): `ctm_cma_avg`, `ctm_nn_avg`, `hopf_avg`, `jr_avg`, `pca_avg`, `psd_ae_avg`, and `wong_wang_avg`.

We follow the same evaluation framework applied throughout this chapter, examining classification accuracy together with mutual information, latent geometry, cluster quality, and dimensionality efficiency. Each representation is paired with the same classifier template, tuned individually, so comparisons reflect the quality of the latent space rather than classifier settings.

Classification Performance

	ctm_cma_avg	ctm_nn_avg	hopf_avg	jr_avg	wong_wang_avg	pca_avg	psd_ae_avg
Input type	Avg psd	Avg psd	Avg psd	Avg psd	Avg psd	Avg psd	Avg psd
Latent size	8	8	16	6	5	8	8
Pre-training	no	yes	no	no	no	yes	yes
Sex accuracy	55,6%	58,2%	**59,6**%	53,0%	55,5%	57,0%	57,0%
Abnormality accuracy	70,0%	74,1%	73,9%	69,6%	61,3%	73,5%	**74,4**%

Table 4.1: Comparison of small-group methods: input types, latent sizes, pre-training use, and accuracies.

On abnormal vs. normal classification, the cortico–thalamic model with amortised inference reaches 74.1% accuracy, close to PCA (73.5%) and the spectral autoencoder (74.4%). Sex classification is more difficult, with best accuracies in the 55–60% range. Two patterns emerge from these results: data-driven methods tend to perform slightly better in classification tasks, while larger latent dimensions do not guarantee improved performance for computational brain models. Since all approaches operate on power spectra, performance is ultimately bounded by how much task-relevant information is present in the PSD itself.

Dimensionality and Efficiency

In this section we want to illuminate the following question. "How efficiently does the method use its representational capacity?" Figure 4.1 Gives us pretty solid insight.

In our implementation we define a dimension as active if it's per-dimension variance is above a threshold of 10^{-3}. There are a couple of things that are interesting about this plot:

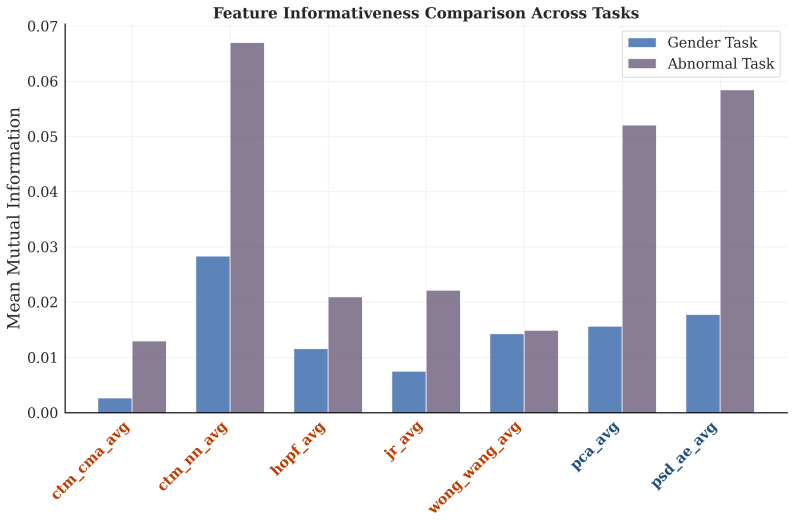


Figure 4.2: Comparison of mean mutual information across all latent dimensions between each Small group representation and the downstream task labels.

- On average, dimensionality efficiency is high, so the extraction methods use most of their available latent capacity.
- The two data-driven methods make use of their latent dimension, as it lies in their nature to do so[9, 36].
- Our hybrid approach, the amortised regression estimator coupled with the cortico-thalamic model, actually has a 100% utilisations of it's dimensions, even though its CMA-ES counterpart does not achieve that.
- While our computational brain models still have high efficiency, besides the CMA-ES cortico-thalamic model, two of them have a dimension which is not active, the reader has to keep in mind that one of them only has 5 dimensions while the other one has 16 is therefore the biggest latent space in our small group.

This is a nice first indication that the computational brain models capture information/variation, however there are still some other factors that contribute to how useful these information are. One of them is the information content in these latent representations.

Information Content

Here we assess how much task-relevant information is encoded in the latent representations by estimating the mutual information between the extracted features and the downstream tasks.

Figure 4.2 shows that, while the sex classification task exhibits substantially lower mutual information overall, the data-driven methods consistently achieve the highest values. Although their advantage in the sex task is relatively modest, they clearly outperform all alternatives in the abnormality detection task. Notably, across both tasks the hybrid approach `ctm_nn_avg` emerges as the best-performing method. To draw general conclusions, a broader range of downstream tasks would be required, yet these results already indicate a consistent trend: certain methods, particularly the hybrid variant, encode considerably more task-relevant information than others.

Feature Independence/Disentanglement

When we talk about feature independence or disentanglement (Figure 4.3) we mostly wonder about the question how well-structured/interpretable the latent features are. We assess feature independence, which can be seen as a proxy for disentanglement, using the Hilbert–Schmidt Independence

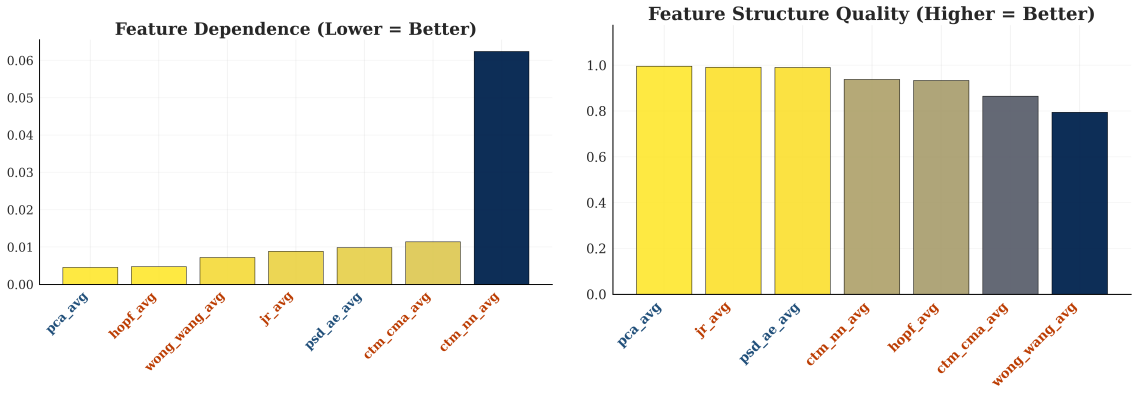


Figure 4.3: Visualisation Feature Dependence and Feature Structure Quality for the Small group methods.

Criterion (HSIC). Lower values indicate greater independence among latent dimensions.

Most methods achieve very low scores in the range of 0.004–0.011, suggesting that they produce largely decorrelated representations. In contrast, the hybrid method `ctm_nn_avg` yields a substantially higher value (around 0.06), indicating that its features are less independent and thus quite different to its CMA-ES counterpart. This result suggests that, while the hybrid approach is competitive in other metrics, it does not promote disentanglement as effectively as the baselines.

As a complementary perspective, we also examine feature structure quality, defined as $(1 - \text{HSIC}) \times \text{Efficiency}$. Here, `pca_avg`, `jr_avg`, and `psd_ae_avg` perform best, attaining values close to the theoretical maximum. The hybrid approaches (`ctm_nn_avg`) and dynamical models (`ctm_cma_avg`, `hopf_avg`, `wong_wang_avg`) score slightly lower but remain competitive.

Overall, these findings reveal a contrast: while the hybrid approach excels in other aspects of evaluation, it provides weaker disentanglement than the baselines, and its feature structure quality does not surpass that of simpler linear or autoencoder methods.

Geometric / Neighborhood Preservation

We next evaluate how well the latent representations preserve the geometric structure of the channel-averaged PSD feature space. This analysis considers both local and global aspects: *trustworthiness* quantifies neighborhood preservation by measuring whether local neighbors in the latent space are also neighbors in the PSD space. *Continuity* assesses embedding smoothness, i.e. whether points that are close in the PSD space remain close after transformation; and *distance correlation* measures global distance preservation across all pairs of samples. To summarize these three aspects, we also report an *overall geometric score*, defined as the composite of the individual metrics (Figure 4.4).

The results show a clear advantage for the data-driven baselines. Both `pca_avg` and `psd_ae_avg` achieve nearly perfect preservation, with trustworthiness and continuity above 0.98 and distance correlations of 0.975 and 0.926, respectively. Their composite scores indicate that they provide almost isometric reductions of the averaged PSD manifold.

Among the CBMs, the hybrid model `ctm_nn_avg` performs best, reaching trustworthiness 0.894, continuity 0.911, distance correlation 0.755, and a composite score of 0.853. This suggests that the hybrid preserves local neighborhoods and global distances to a substantial degree, although not as completely as PCA or the PSD autoencoder. Other CBMs perform considerably worse: `ctm_cma_avg`, `jr_avg`, `hopf_avg`, and `wong_wang_avg` yield composite scores between 0.41 and 0.47, driven by weak neighborhood preservation and near-zero or negative distance correlations.

This indicates that while the hybrid CTM achieves meaningful geometric preservation, purely mechanistic embeddings tend to distort PSD geometry strongly. In contrast, data-driven ap-

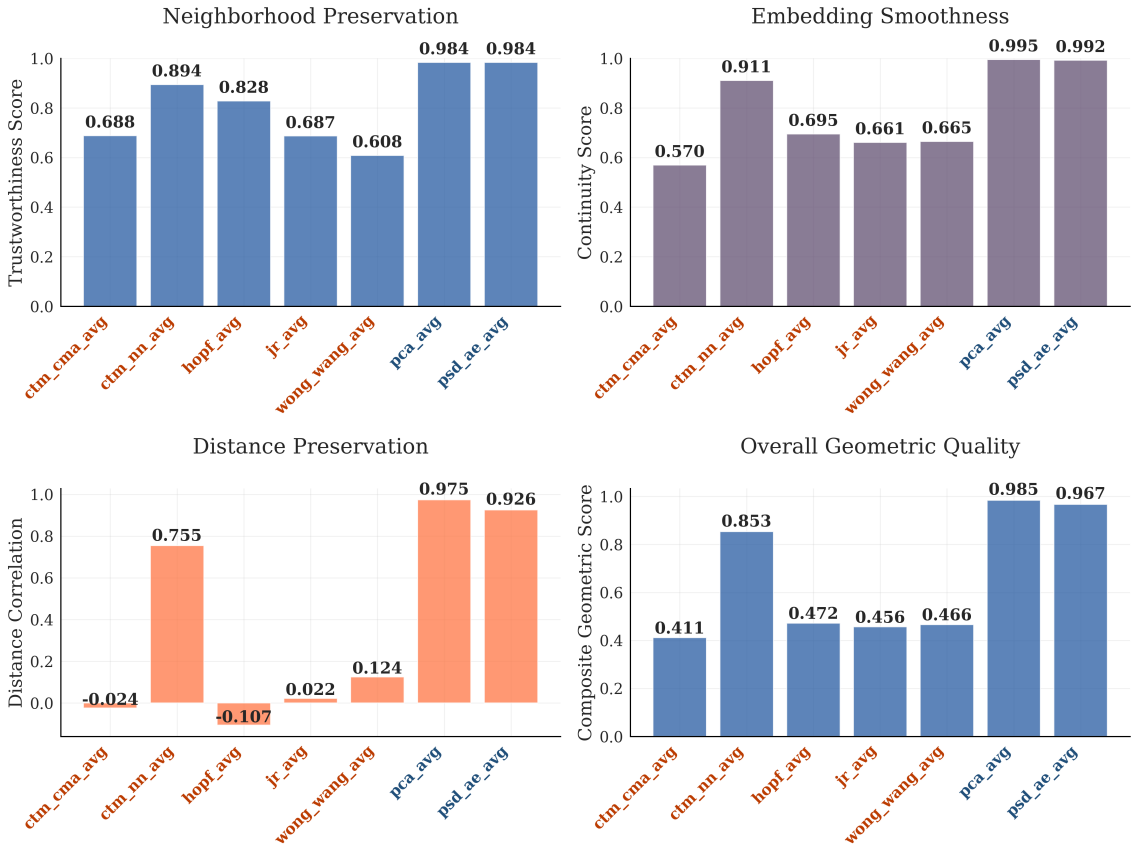


Figure 4.4: Trustworthiness, continuity, and distance correlation scores for each Small group method with respect to averaged PSD features.

proaches such as PCA and PSD–AE set the upper bound for geometric fidelity. The Discussion in Chapter 5 considers what this trade-off between mechanistic interpretability and geometric accuracy implies for the practical usefulness of CBMs.

Cluster Quality

To assess how well the extracted representations separate similar from dissimilar samples, we evaluated cluster quality using three complementary metrics: the *Silhouette score*, the *Davies–Bouldin (DB) index*, and the *Calinski–Harabasz (CH) index* (Figure 4.5).

The Silhouette score measures average separation between clusters, with higher values indicating better-defined cluster boundaries. The DB index evaluates compactness and separation simultaneously, where lower values are better. The CH index reflects the variance ratio between clusters and within clusters, rewarding high between-cluster separation and low within-cluster scatter.

Across all metrics, `wong_wang_avg` achieved the strongest clustering performance, with the highest Silhouette score (0.636), the lowest DB index (0.48), and by far the largest CH index (55,898). The hybrid approach `ctm_nn_avg` also performed competitively, with a Silhouette score of 0.511, a DB index of 0.54, and a CH index of 14,029, outperforming most other baselines. The `ctm_cma_avg` and `hopf_avg` models yielded moderate scores across all measures, whereas linear and autoencoder-based methods (`jr_avg`, `pca_avg`, and `psd_ae_avg`) consistently underperformed, showing low Silhouette scores (< 0.31), high DB indices (> 1.0), and relatively small CH indices ($< 3{,}000$).

The composite clustering score, which normalizes and combines all three metrics, confirms these trends: `wong_wang_avg` clearly dominates (0.766), followed by `ctm_nn_avg` (0.462). Other methods cluster tightly around intermediate values (e.g., `ctm_cma_avg` 0.374, `hopf_avg` 0.376), while

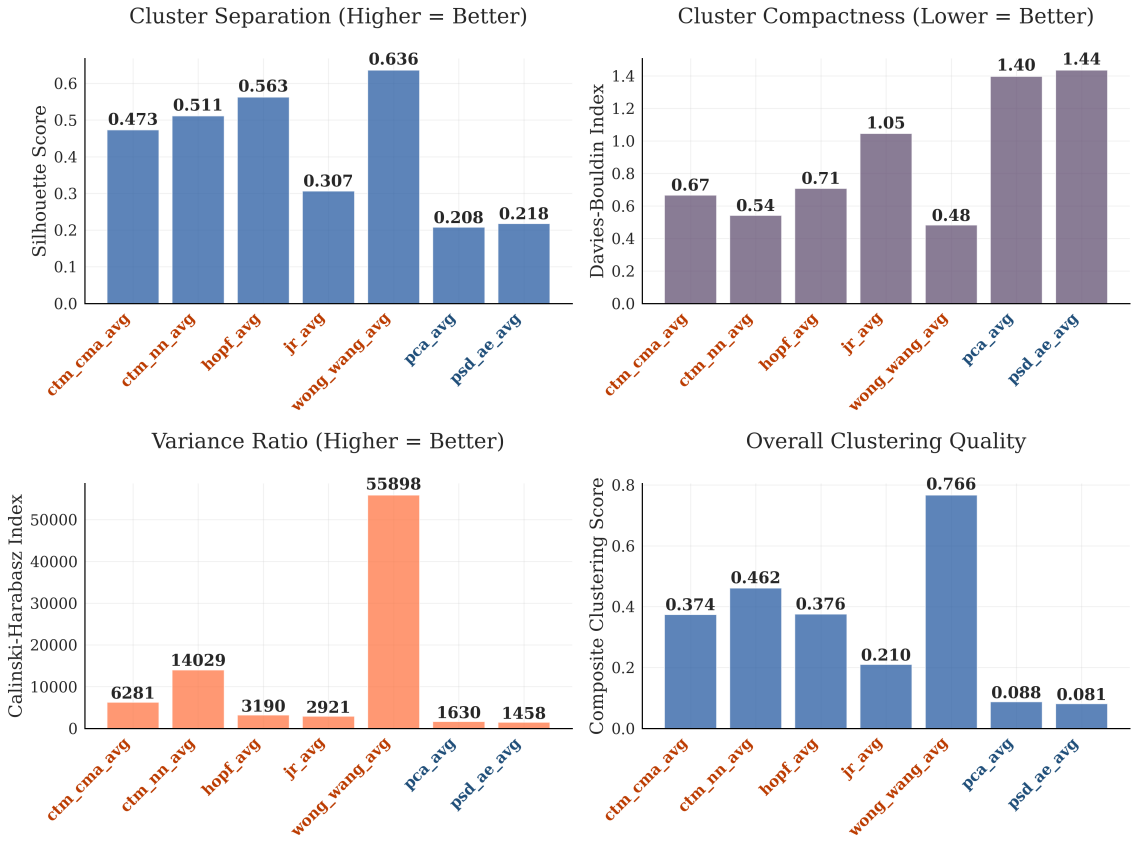


Figure 4.5: Silhouette Score, Davies-Bouldin and Calinski–Harabasz index for each Small group method.

linear baselines remain low (< 0.21).

It is worth noting that the Wong–Wang and JR models operate in smaller latent dimensionalities (5 and 6, respectively) compared to most other methods (8 dimensions) and especially the Hopf model (16 dimensions). Lower latent dimensionality can reduce within-cluster scatter and thus improve clustering scores, particularly the Calinski–Harabasz index. While this effect likely contributes to Wong–Wang's strong clustering results, the much weaker performance of JR despite a comparable dimensionality indicates that methodological differences also play a critical role.

Overall, these results indicate that the dynamical `wong_wang_avg` model provides the most coherent clustering structure, with the hybrid `ctm_nn_avg` also offering a substantial improvement over either data-driven methods or its CMA-ES related variant. This suggests that models incorporating neural and biophysical dynamics can produce representations that organize similar data points more effectively than purely linear or autoencoder-based approaches.

Representation Alignment/Similarity

To see how similarly the methods organize information in latent space, we compare them with several pairwise similarity measures (Figure 4.6). *CCA max-correlation* and *linear CKA* capture linear subspace alignment (invariant to rotation and uniform scaling), *RBF CKA* is sensitive to nonlinear correspondences, *distance–geometry correlation* (Pearson correlation of pairwise Euclidean distances) reflects agreement of global geometry, *KNN Jaccard* (k=10) measures overlap of local neighborhoods, and *Procrustes disparity* (lower is better) quantifies residual mismatch after optimal orthogonal alignment.

A clear pattern emerges. The data-driven baselines together with the hybrid CTM (`pca_avg`, `psd_ae_avg`, `ctm_nn_avg`) align strongly under CCA/linear CKA and show low Procrustes disparity, i.e., they span closely related subspaces. Within this trio, `pca_avg` and `psd_ae_avg` also

share substantial local structure (high KNN overlap), whereas `ctm_nn_avg`, despite strong linear alignment,differs more in global distance geometry, suggesting different variance weighting across axes. Apart from a relatively high CCA alignment with `hopf_avg`, these three methods are generally more similar to one another than to the computational brain models.

Among the dynamical models, `wong_wang_avg` is the most distinct: it shows weaker CKA/CCA alignment to other CBMs, only slightly higher distance–geometry agreement with the data-driven group, and a larger Procrustes disparity. In other words, its representation is not just a rotated or rescaled version of the others, a finding that fits with its strong clustering (Figure 4.5) and high geometric preservation (Figure 4.4). `hopf_avg` sits between groups, sharing some linear alignment but differing under nonlinear CKA and in local neighborhoods.

Taken together, these similarity results complement the geometry and clustering analyses: methods can preserve global geometry (Figure 4.4) yet realize different local neighborhoods, and strong clustering (Figure 4.5) does not require close linear alignment with other representations. Thus, representational *quality* and *congruence across methods* are related but distinct.

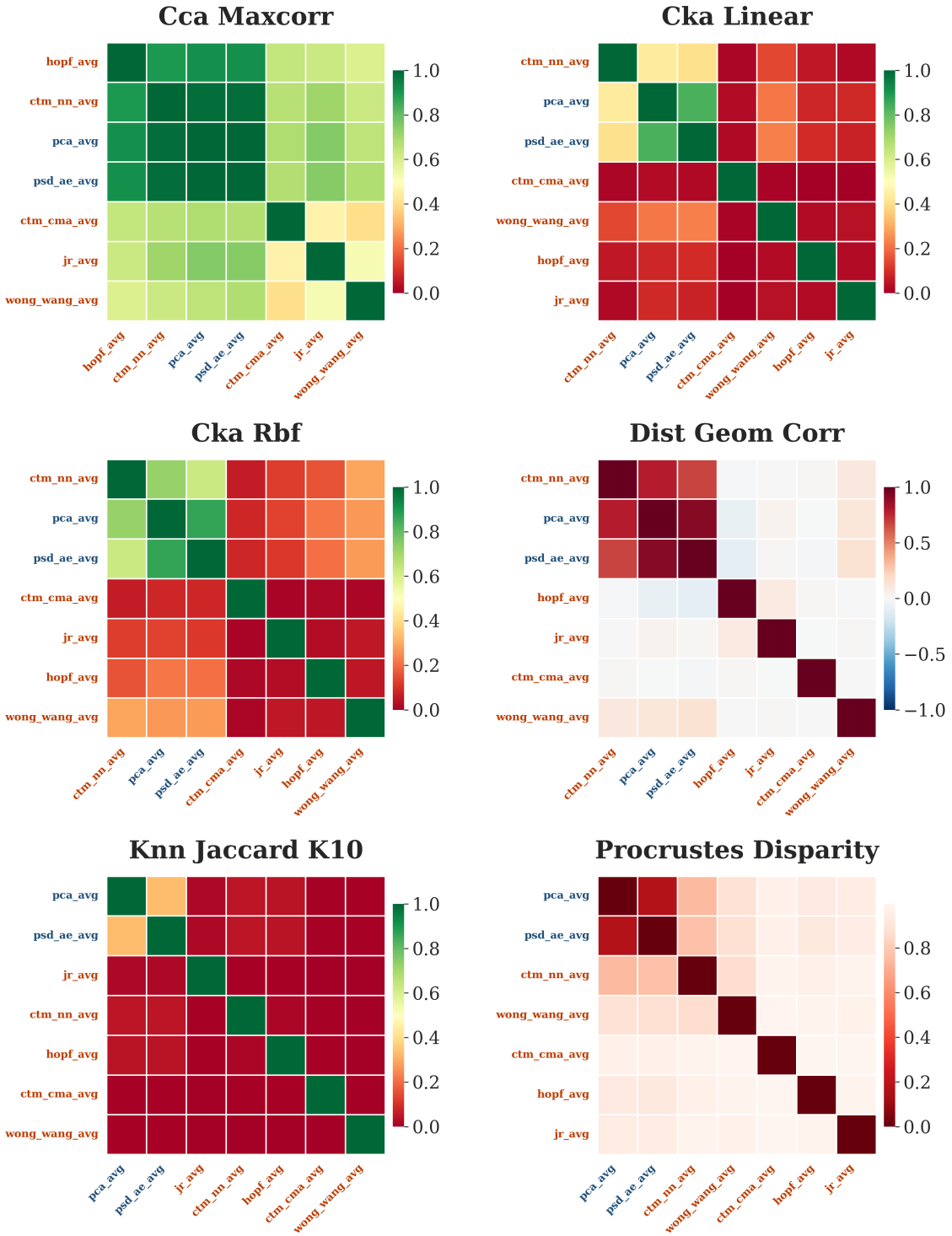


Figure 4.6: Pairwise representation similarity across methods for the CBMs and data-driven methods in the Small group. Higher values indicate stronger agreement for CCA/CKA, distance–geometry correlation, and KNN Jaccard; lower values indicate better alignment for Procrustes disparity.

4.2 Weight Class - Medium

For the *Medium* weight class, we evaluate a very similar set of representation methods under moderately stronger connection weights. Namely: ctm_nn_pc, hopf_pc, jr_pc, c22, pca_pc, psd_ae_pc and eegnet.

Following the multi-faceted evaluation framework from the *Small* weight class, we examine classification performance alongside a range of representation quality metrics. We again emphasize that all methods are paired with the same logic, but independently tuned classifier so each methods has the best chance of achieving performance. The motivation came from the observation that we have vastly different latent sizes. This ensures that any performance gains in the Medium condition stem from improved representations rather than just fortunate or unfortunate classifier configurations.

Classification Performance

	ctm_nn_pc	hopf_pc	jr_pc	catch22	pca_pc	psd_ae_pc	eeg_net
Input type	Per-chan psd	Per-chan psd	Per-chan psd	Per-chan empirical	Per-chan psd	Per-chan psd	Per-chan empirical
Latent size	$8 \times 19 = 152$	$16 \times 19 = 304$	$6 \times 19 = 114$	$22 \times 19 = 418$	$8 \times 19 = 152$	$8 \times 19 = 152$	128
Pre-training	yes	no	no	no	yes	yes	yes
Sex accuracy	59,2%	60,7%	57,3%	**62,0**%	60,5%	57,7%	58,5%
Abnormality accuracy	78,4%	76,8%	72,2%	**79,3**%	77,2%	78,2%	78,3%

Table 4.2: Comparison of medium-group methods: input types, latent sizes, pre-training use, and accuracies.

Table 4.2 summarizes the classification results for the Medium weight class. Compared to their Small-class counterparts, all methods achieve higher accuracies, which can be attributed not only to the larger latent dimensionality but also to the use of per-channel inputs rather than aggregated spectra. Access to individual channel power spectra provides additional task-relevant variability, which appears particularly useful for abnormality detection.

Among the methods, `catch22` achieves the best overall performance, reaching 62.0% on the sex task and 79.3% on the abnormality task. The hybrid `ctm_nn_pc`, `psd_ae_pc`, and `eegnet` follow closely, all attaining abnormality accuracies above 78%. Domain-driven representations such as `hopf_pc` and `jr_pc` perform more modestly, with `jr_pc` yielding the weakest results across both tasks (57.3% and 72.2%). Linear PCA features (`pca_pc`) fall in the lower-middle range: stronger than `jr_pc` but behind the best-performing approaches.

The variance in performance across methods is greater here than in the Small class. In particular, methods that can exploit richer per-channel structure (e.g., `catch22`, `eegnet`) pull slightly, suggesting that finer-grained input representations allow advanced models to capitalize on class-discriminative patterns inaccessible to simpler linear or mechanistic baselines. It should be noted, however, that `catch22` benefits from a substantially larger latent dimensionality, which may contribute to its advantage.

Overall, even though the strongest (and largest) method in this group is data-driven, most of the computational brain models deliver performance that is comparable to the data-driven approaches across both tasks.

Dimensionality and Efficiency

Figure 4.7 shows the distribution of active versus inactive feature dimensions across methods. As before, we define a dimension as *active* if it carries non-negligible variance, while *inactive* dimensions contribute little and can be considered unused capacity. In the Medium weight class, methods vary in how effectively they exploit their representational space. Approaches such as `ctm_nn_pc` and `psd_ae_pc` activate nearly all dimensions, indicating that they distribute information broadly

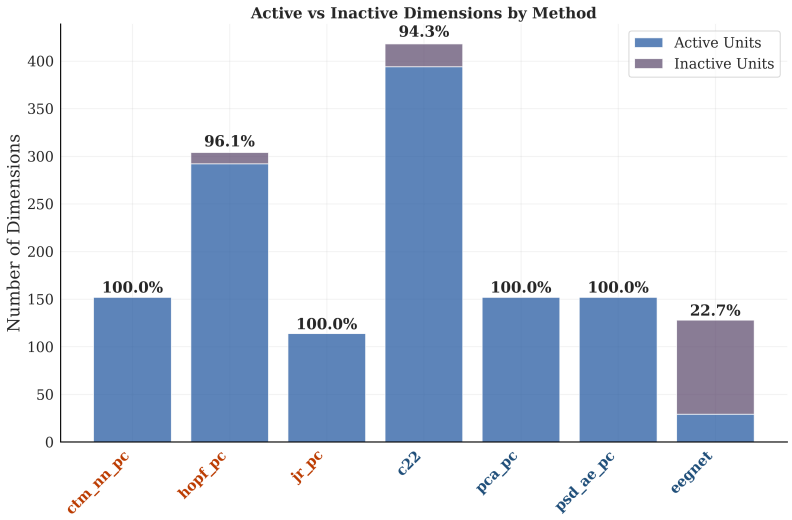


Figure 4.7: Dimensionality efficiency for the Medium group methods. Each bar shows the number of active and inactive latent dimensions per method, with dimensions flagged as inactive if they explain less than 10^{-3} variance. This highlights differences in how efficiently each method utilizes its latent capacity.

across their feature set. By contrast, `eegnet` and `Catch22` exhibit a substantial proportion of inactive dimensions, suggesting a lower intrinsic dimensionality in their learned representations. In the case of `Catch22`, this limitation likely arises from the fixed nature of its handcrafted feature set, whereas `eegnet` may actively condense most relevant information into a smaller subspace than provided. This behavior is consistent with its training objective: the PSD-based reconstruction loss requires the network to capture the overall spectral structure of the EEG signals, but not to utilize every available latent unit. Compared to the Small weight class, the overall efficiency of latent utilisation remains similar for the shared methods, with the computational brain models again showing near-complete activation of their available dimensions, as expected given that the Medium group simply scales their per-channel representations.

When diving a little bit deeper into the latent space analysis of `eegnet` by looking at the figures 4.8 and 4.9. We can see that the reason for its poor efficiency becomes clear. Out of 128 available dimensions, only about 29 (22.7%) are consistently active across both training and evaluation. The remaining 99 dimensions carry virtually no variance (between 10^-4 and 10^-3) and are therefore unused capacity. While the features that `eegnet` does use may be meaningful, the vast majority of its latent units remain inactive, leading to an inefficient representation.

This indicates that the poor efficiency is not due to instability or overfitting, but rather to a mismatch between model capacity and training objective. The PSD-based reconstruction loss is comparatively simple, and `eegnet` might possess more representational power than is required to optimize it. As a result, the network theoretically could minimize the loss by concentrating information into a small subset of dimensions while leaving the majority inactive. In this sense, the inefficiency arises not from a failure of the model to learn, but from an overparameterisation relative to the simplicity of the objective. However, we can't say for sure, until we would've tested different configurations of the EEGNet.

Information Content

We quantify each feature's class informativeness by computing the mutual information (MI) between that feature dimension and the class label (Figure 4.10). Overall, the Medium weight class yields slightly higher MI values compared to the Small class, reflecting that stronger connectivity weights amplify class-specific structure in the data. Despite this general increase, the relative ordering of methods remains stable. The hybrid approach `ctm_nn_pc` and the handcrafted `Catch22` features encode the largest amount of task-relevant information, clearly outperforming both data-driven baselines and dynamical models. PCA and the PSD autoencoder achieve intermediate

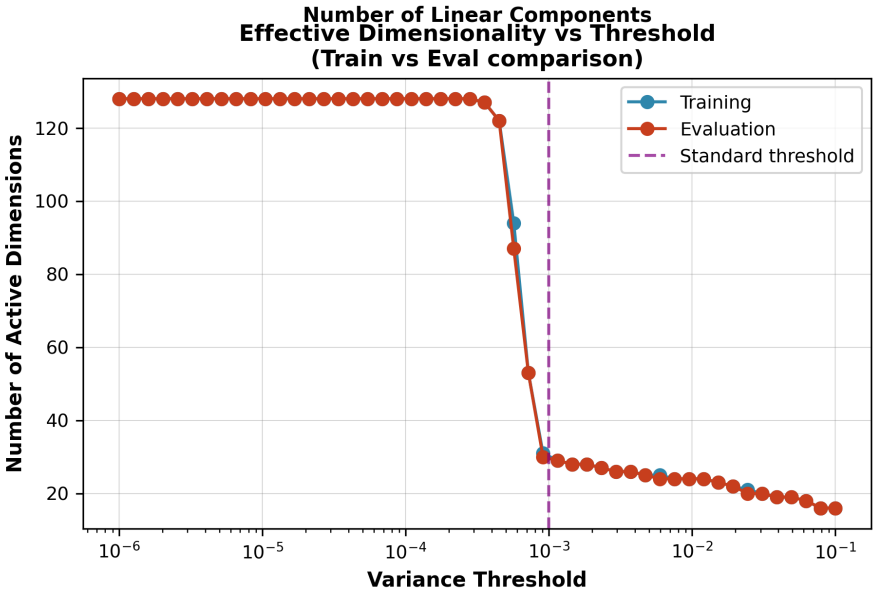


Figure 4.8: Active Dimensions vs. Variance Threshold of EEGNet.

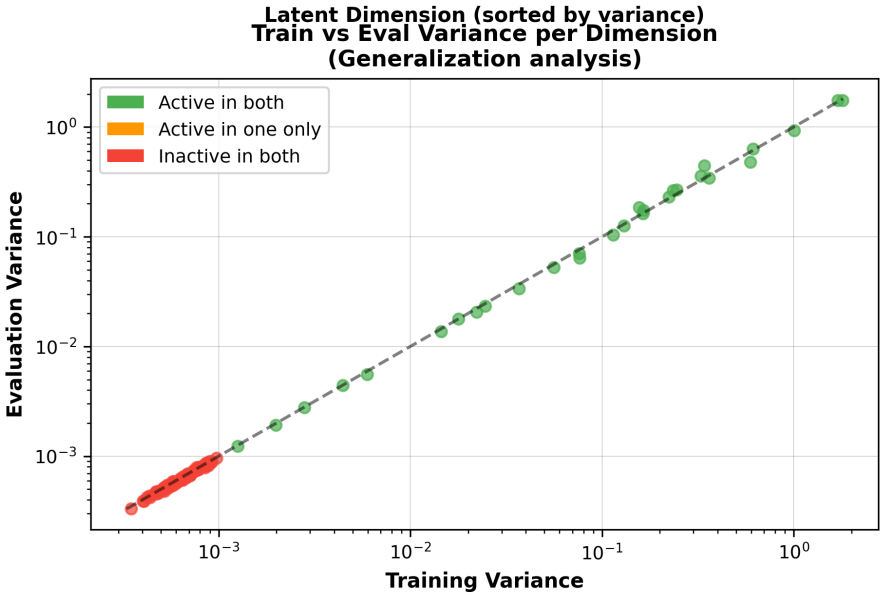


Figure 4.9: Train and Evaluation Latent Dimension Variance of EEGNet.

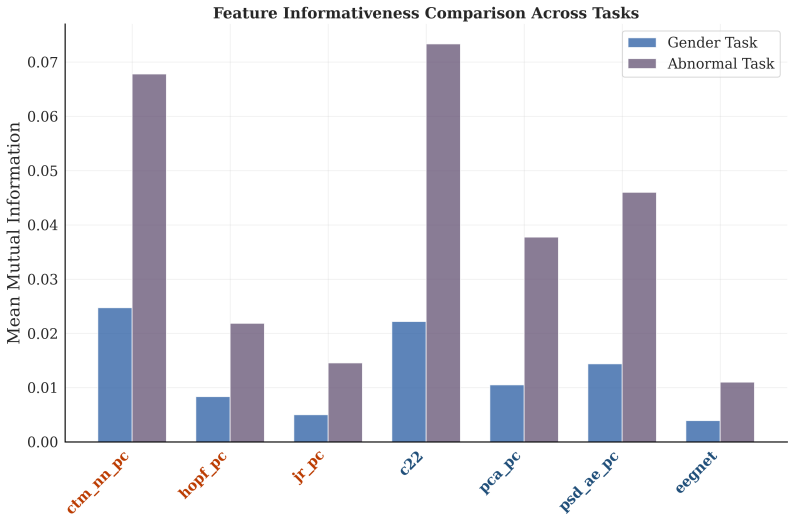


Figure 4.10: Comparison of mean mutual information across all latent dimensions between each "Medium" representation and the downstream task labels.

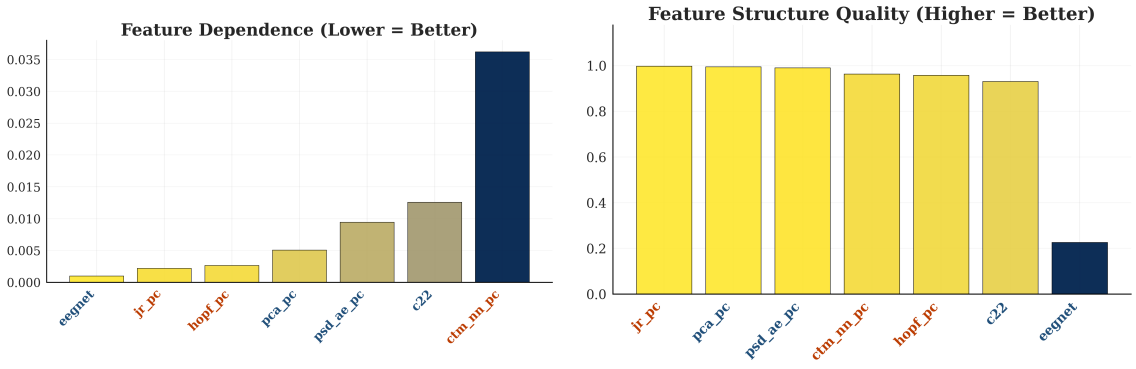


Figure 4.11: Visualisation Feature Dependence and Feature Structure for the Medium group methods.

scores, showing that variance-preserving or reconstruction-oriented methods can capture some, but not all, of the discriminative signal. By contrast, the dynamical models (`hopf_pc`, `jr_pc`) continue to encode only limited information, suggesting that while they preserve structural properties, they are less aligned with the downstream tasks. Most striking is the consistently low MI of `eegnet`, which lags behind all other methods. This likely reflects the fact that the network compresses relevant information into only a subset of dimensions, driven by the PSD-based reconstruction loss, rather than distributing it broadly across the latent space. Taken together, these results emphasize that in the Medium regime, methods with explicit task-relevant inductive biases (hybrid or handcrafted features) capture the richest information, whereas purely dynamical or heavily compressed representations remain comparatively weak. The pattern mirrors the Small class but with sharper contrasts, highlighting that increased signal strength accentuates differences between representation families.

Feature Independence/Disentanglement

Beyond informativeness, an effective representation should encode information with minimal redundancy across latent dimensions. We measure this using the Hilbert–Schmidt Independence Criterion (HSIC), averaged across all feature pairs, where lower values indicate greater independence (Figure 4.11, left).

The results reveal clear differences between methods. `eegnet` achieves the lowest dependence values, indicating that its latent features are highly decorrelated from one another despite its large

dimensionality. The biophysical models `jr_pc` and `hopf_pc` also produce largely independent features, reaching levels of independence comparable to PCA (even better). By contrast, the PSD autoencoder and `Catch22` exhibit moderately higher dependence, which is plausible given that they often encode overlapping frequency components or statistical descriptors across channels. The highest dependence is observed for `ctm_nn_pc`, whose features are substantially more correlated, suggesting a more entangled representation in which information is distributed redundantly across dimensions.

To contextualize this, we also consider overall feature structure quality, which combines independence with efficiency of dimension usage (Figure 4.11, right). Here, almost all methods except `eegnet` achieve values close to the theoretical maximum, demonstrating that even when some correlations exist, the features remain well-structured and make effective use of available capacity. The strong penalisation of `eegnet` arises because it activates only a small portion of its latent dimensions (cf. Section 4.2), which substantially reduces its efficiency score despite producing decorrelated features. By contrast, the high structure quality of PCA, `jr_pc`, and `psd_ae_pc` indicates that these methods achieve a good balance: they use nearly all of their latent dimensions while keeping redundancy low.

Taken together, these findings highlight a trade-off. Methods like `ctm_nn_pc` utilize their full latent space but at the cost of stronger inter-feature dependencies, while methods such as `eegnet` achieve independence but waste large portions of their representational capacity. The most effective compromises appear in PCA, the PSD autoencoder, and the simpler biophysical models, which preserve both independence and structural efficiency.

Geometric / Neighborhood Preservation

We next assess how well each representation preserves the geometric relationships of the PSD feature space, using trustworthiness, continuity, and distance correlation as complementary measures (Figure 4.12).

Trustworthiness evaluates whether local neighbors in the PSD space remain close in the latent, while continuity checks the converse: whether latent neighbors are also neighbors in the PSD space. Distance correlation captures global fidelity by comparing pairwise distances across all samples. These three scores are also combined into a composite index of overall geometric quality.

The strongest overall performance is achieved by the data-driven baselines. PCA and the PSD autoencoder obtain the highest composite scores (0.969 and 0.950), with trustworthiness around 0.96 and continuity near 0.98. Their distance correlations (0.958 and 0.902) confirm that they preserve both local neighborhoods and global geometry very closely.

Among the CBMs, the hybrid `ctm_nn_pc` stands out with a composite score of 0.875, supported by trustworthiness 0.885, continuity 0.934, and distance correlation 0.805. This shows it can maintain a substantial portion of the PSD manifold structure, though not as completely as PCA or PSD–AE. By contrast, the `hopf_pc` and `jr_pc` embeddings perform considerably worse, with composite scores of 0.297 and 0.439, reflecting weak neighborhood preservation and very low distance correlations (negative for Hopf).

The empirical feature baseline `Catch22` achieves a modest composite score (0.461): its trustworthiness is better than Hopf or JR, but it lacks global distance preservation. EEGNet sits in between these extremes, with good trustworthiness (0.823) and moderate continuity (0.795), but a low distance correlation (0.110) that limits its composite score (0.576).

In summary, the medium group confirms the pattern seen for the Small group: PCA and PSD–AE provide nearly isometric reductions of PSD features, the hybrid CBM preserves geometry better than other mechanistic models, while purely dynamical CBMs sacrifice geometric fidelity. EEGNet retains local neighborhood structure but distorts global relationships, consistent with its emphasis on discriminative rather than geometry-preserving embeddings.

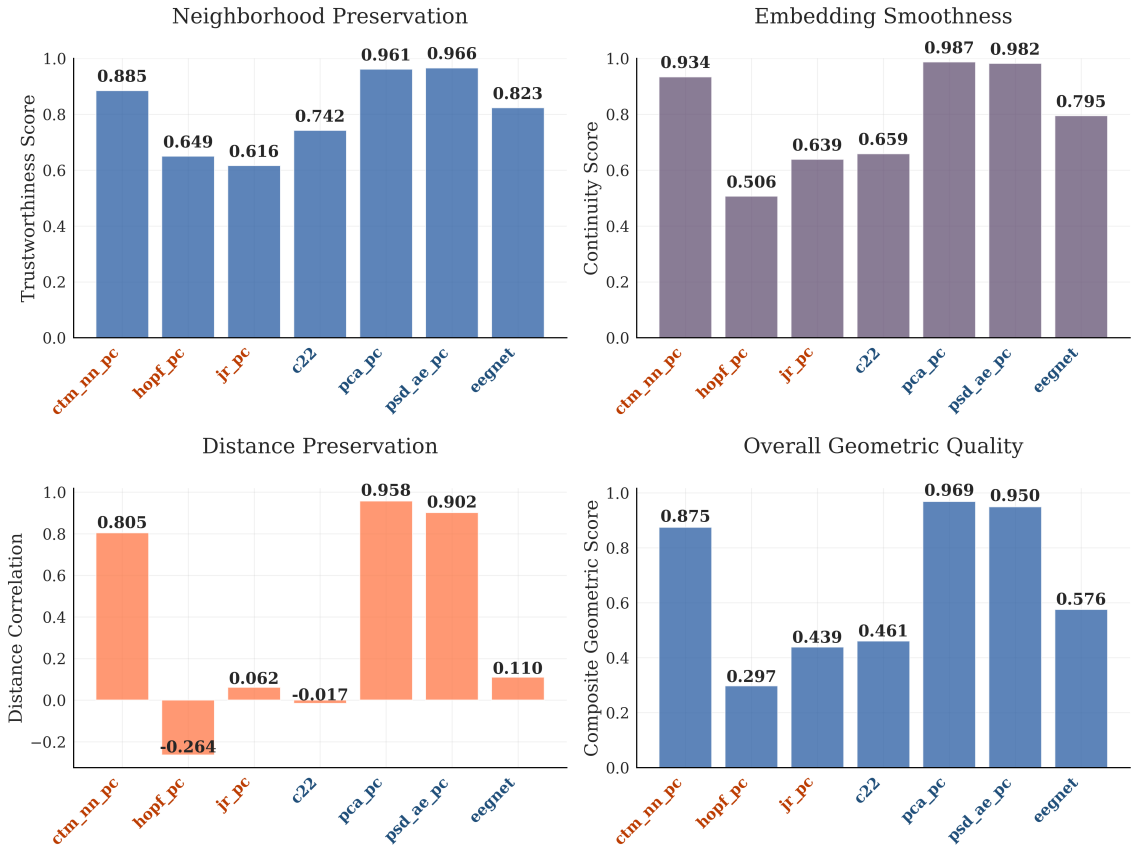


Figure 4.12: Trustworthiness, continuity, and distance correlation scores for each Medium group method with respect to per-channel PSD features.

Cluster Quality

To evaluate whether the representations naturally separate samples into class-like groups without supervision, we examined clustering quality using the Silhouette score, Davies–Bouldin index, and Calinski–Harabasz index, along with a composite score (Figure 4.13). These metrics quantify how compact and well-separated the unsupervised clusters are. Higher values indicate that the latent space organizes data in a way that is more consistent with the underlying class structure, even though the class labels themselves are not used in the computation.

The results show a clear stratification among methods. The hybrid model `ctm_nn_pc` achieves the strongest clustering performance, followed by the handcrafted `Catch22` features. Both methods form relatively compact and well-isolated clusters, consistent with their high mutual information scores reported earlier: representations that encode more task-relevant information also yield clearer unsupervised class structure. EEGNet ranks only in the middle tier. Its clusters are somewhat distinct but less separated, reflecting the earlier observation that it compresses information into a limited subset of features and distorts global geometry. Linear and reconstruction-based methods (PCA, PSD–AE), but especially the dynamical models (Hopf, JR) show weaker clustering, with overlapping or diffuse cluster structure that aligns with their lower informativeness and reduced geometric preservation.

Taken together, these results reinforce a theme observed across evaluation axes: methods that balance inductive bias with task-relevant information capture, such as the hybrid `ctm_nn_pc` and handcrafted `Catch22`, provide the most useful latent organisations. This aligns with their accuracy in downstream tasks. Here, we explicitly want to remind the reader that these are only binary classification and not regression tasks. By contrast, methods that either overcompress (EEGNet) or emphasize generative/variance-preserving criteria (PCA, AE, dynamical models) do not achieve the same degree of separability. Compared to the Small class, the Medium class amplifies these differences: the increased representational capacity sharpens class boundaries in the latent space,

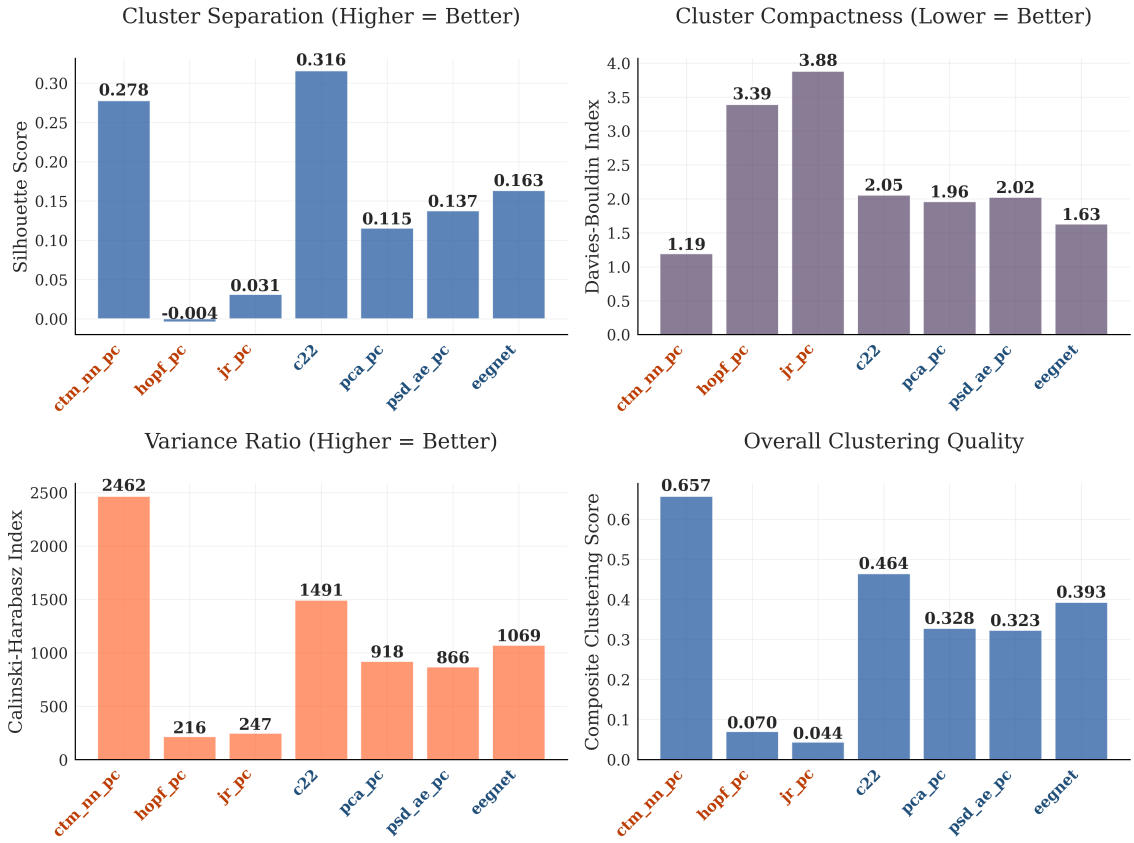


Figure 4.13: Silhouette Score, Davies-Bouldin and Calinski–Harabasz index for each Medium group method.

making the relative strengths and weaknesses of each method more pronounced.

Representational Similarity Across Methods.

We conclude by comparing how closely the different methods align in their latent organisations, using multiple complementary similarity measures (Figure 4.14). CCA max-correlation assesses the maximal linear alignment between subspaces, linear CKA evaluates covariance structure overlap, RBF CKA extends this to nonlinear neighborhood geometry via Gaussian kernels, distance–geometry correlation captures correspondence of global pairwise distances, KNN Jaccard measures local neighborhood overlap, and Procrustes disparity quantifies residual mismatch after optimal rigid alignment.

The results reveal both broad families of methods and notable outliers. Under CCA max-correlation, nearly all methods show strong alignment, with the only pronounced divergence appearing between `hopf_pc`, `ctm_nn_pc` and `Catch22`. Moving to linear CKA, however, the picture becomes sharper: the data-driven approaches (`pca_pc`, `psd_ae_pc`) align moderately with `eegnet`, but both biophysical models (`jr_pc`, `hopf_pc`) fall to the lowest alignment values and do not even align strongly with one another. CKA RBF accentuates this clustering: `pca_pc`, `psd_ae_pc`, and `ctm_nn_pc` group closely, with `Catch22` and `eegnet` occupying an intermediate position, while the dynamical models remain most distinct. Distance correlation, KNN overlap, and Procrustes disparity all confirm this pattern: `pca_pc`, `psd_ae_pc`, and `ctm_nn_pc` form a tight cluster with near-perfect mutual alignment, while `jr_pc`, `hopf_pc`, and `Catch22` diverge strongly, and `eegnet` sits in between, partially overlapping with the data-driven cluster but not matching it completely.

Taken together, these findings highlight three important points. First, PCA and the PSD autoencoder consistently emerge as the closest pair, effectively capturing the same variance structure. Second, the hybrid `ctm_nn_pc` joins this data-driven family under nonlinear and distance-based

measures, even though it diverges from Hopf in pure subspace correlation. Third, EEGNet and Catch22 both occupy intermediate but distinct positions: EEGNet shows some similarity to the data-driven cluster but maintains a unique geometry, consistent with its moderate clustering quality and low efficiency. Catch22, on the other hand, remains consistently weakly aligned, reflecting its handcrafted, channel-wise feature extraction. By contrast, the dynamical CBMs (`hopf_pc`, `jr_pc`) are the most isolated, rarely aligning with one another and diverging from all other methods.

Compared to the Small class, the Medium class amplifies these divisions: stronger signals and higher representational capacity reveal sharper splits into representational families. The data-driven and hybrid methods converge on a shared latent geometry, while the biophysical models separate into a distinct regime, and EEGNet and Catch22 highlight alternative organisational principles. This reinforces a broader theme across our analyses: the choice of representation method yields increasingly divergent "views" of the same data, each emphasising different aspects of underlying neural structure.

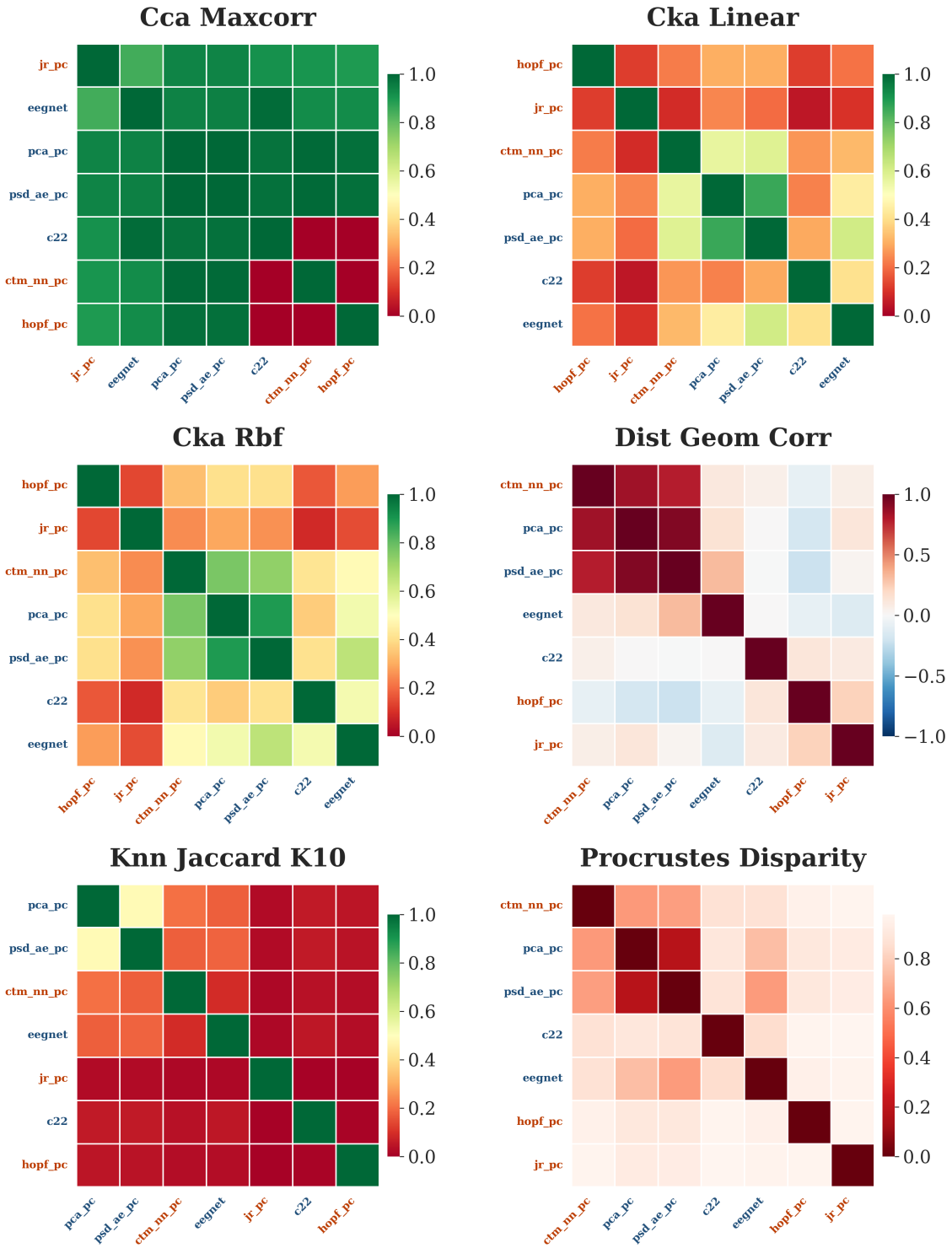


Figure 4.14: Pairwise representation similarity across methods for the CBMs and data-driven methods in the Medium group. Higher values indicate stronger agreement for CCA/CKA, distance–geometry correlation, and KNN Jaccard; lower values indicate better alignment for Procrustes disparity.

Chapter 5

Discussion

Before turning to the structure of the latent spaces, it is useful to briefly reflect on classification accuracy. Across both, the Small and Medium, settings we found that the different approaches reached broadly comparable levels of performance. No single family of methods was clearly dominant, although hybrids such as the cortico–thalamic network (amortised) and feature sets like `catch22` often matched or slightly surpassed the other extraction methods. Pure mechanistic models trailed somewhat in predictive power, yet still delivered accuracies relatively close to PCA and autoencoder methods.

An important detail is how this accuracy was achieved relative to the number of latent dimensions available. Some models reached competitive scores with only a handful of parameters, while others relied on much larger latent spaces. EEGNet is a striking case: despite having more than a hundred potential units, it concentrated class-relevant information into only a small active subset. In contrast, while `catch22` achieved the highest accuracy, it also operated with the largest latent dimension, and part of its strength comes from leveraging this broad, diverse feature set. This means that differences in raw accuracy do not directly reflect differences in representational strength, since some methods achieve similar predictive power with far more compact or efficient codes.

Because the overall performance gap is modest, the more meaningful distinctions emerge when we examine how the latent spaces are organized. Whether parameters are used efficiently, how much class information is encoded, how independent the features are, and how well geometric structure is preserved. At the same time, the performance of CBMs is not fixed but depends on how their parameters are inferred. In our experiments, evolutionary strategies and amortised inference produced somewhat different outcomes, highlighting that optimisation choices can shape the resulting representations. Thus, while accuracies across methods remain close, the latent space analyses provide a clearer view of their respective strengths and trade-offs.

5.1 CBMs use parameters efficiently; autoencoders underutilize

Across both weight classes, we find that most methods, including the CBMs, make near-complete use of their available latent dimensions. In the Small class (5–16 latent features), all methods utilize the bulk of their representational capacity, with only one or two dimensions flagged as "inactive" (near-zero variance). The trend persists in the Medium class, where increased latent size and per-channel inputs give models more room to expand. Notably, the **CBMs consistently exhibit high dimensional efficiency**: nearly every parameter dimension varies meaningfully, reflecting the fact that each corresponds to a biophysical degree of freedom that the model must adjust to fit the data. The hybrid `ctm_nn` in particular stands out for 100% utilisation in the Small class, and it remains among the top in the Medium class.

In contrast, a few data-driven methods reveal **underutilised capacity**. The `eegnet` autoencoder is the clearest example, despite its large latent size (128), it effectively uses only ∼23% of those units, compressing most information into a much smaller subspace and leaving the majority of features virtually dormant (as confirmed by variance analysis in Fig. 4.7). This inefficiency

likely stems from **overcapacity relative to its training objective**: since EEGNet was trained to reconstruct the power spectrum, it can achieve a low reconstruction error by encoding the most salient spectral patterns into a limited set of filters, without needing to activate every latent unit. A similar but less extreme pattern is seen with the `catch22` feature set, which has a fixed 22 features per channel. Here some dimensions carry redundant or negligible variance, implying that not all of the handcrafted descriptors are informative for the given data, but considering the fact that it's non dynamical, we still get high efficiency (see Figure 4.7).

By comparison, methods like PCA and the PSD autoencoder naturally span all available components (PCA by definition uses orthogonal axes of maximal variance, and the autoencoder was constrained to 8 features per channel, which it trained to make all useful). The **mechanistic models** (CTM, JR, Wong–Wang, Hopf) are likewise near-fully efficient. Each parameter in these models influences the EEG power spectrum in a distinct way, so the fitting process tends to adjust all parameters rather than leaving any consistently at default.

In summary, **CBMs prove to be very parameter-efficient feature extractors**: even with small latent sizes they capture most variance available, and when scaled to per-channel representations they naturally continue to use each dimension (due to our setup, see chapter 3). Data-driven methods can also be efficient, but those with unconstrained high capacity may require careful regularisation or tuning to avoid unused dimensions. This high utilisation by CBMs bodes well for their practicality. It indicates that the interpretable parameters are not, for the most part, "dead dimensions" but carry signal, an encouraging sign that the models are capturing a broad range of EEG variance rather than collapsing onto a few factors. It also highlights a nuanced difference in **inductive bias**: a mechanistic model, with each dimension tied to a physiological quantity, inherently pressures the representation to make use of all those quantities. A deep autoencoder, in contrast, is free to ignore any neurons that do not help reduce the loss. Thus, **in terms of raw dimensional efficiency, CBMs behave similarly to standard methods (often near the theoretical maximum), and can even surpass overly flexible models that may not fully utilize their latent space**. This efficiency, however, does not alone guarantee that the representation is *useful*. For that, we must examine whether the information they carry is useful and how they organise the data.

5.2 Data-driven features encode more class information

Our analysis of mutual information (MI) between latent features and class labels shows a consistent advantage for learned or hand-crafted representations compared to purely mechanistic ones. In the Small weight class, PCA and PSD–AE tend to rank higher on MI for abnormal versus normal EEG, and they retain a slight edge on the subtler sex task. The hybrid CTM (`ctm_nn`) stands out within the CBM family, ranking at or near the top across both tasks and representing the strongest CBM overall. Although its encoder is trained in an unsupervised way, it appears to extract feature directions that align with class structure in the data. By contrast, the pure CBMs such as Jansen–Rit and Hopf carry less label information on average, consistent with their more constrained parameterisations.

A likely contributor to this gap is misspecification in the higher-frequency range. In the fit-quality analysis, most CBMs underfit the broad "beta bump" above about 15 Hz, leaving systematic residuals in the 15–25 Hz band (see 3.4.10 and Appendix A). This behaviour is especially marked for Wong–Wang, where cases with negative R^2 highlight a mismatch between model and data at higher frequencies. CMA–ES fits of the CTM generally follow the dominant alpha and beta peaks more closely than the amortised variant, although both exhibit some high-frequency residuals. If discriminative variance is concentrated in this range, systematic underfit will reduce MI. Despite sharing the same residual structure, the amortised CTM remains competitive, indicating that other aspects of its latent representation compensate for this missing spectral feature.

In the Medium class, where inputs are per-channel and latents are larger, MI values increase modestly across all methods but the relative ordering remains stable. Catch22 features and the hybrid CTM occupy the upper tier, while linear baselines sit in the middle and dynamical CBMs trail. Channel-specific detail appears to benefit methods that can flexibly exploit spectral differences, while fixed-form priors are less able to capitalise on this additional information. The EEGNet autoencoder continues to encode surprisingly little class information despite its capacity,

underscoring that larger latent spaces do not guarantee discriminative utility.

Differences in information content are reflected in the intrinsic clustering structure of latent spaces. In the Small class, Wong–Wang achieves the strongest cluster validity across Silhouette, Davies–Bouldin, and Calinski–Harabasz indices, followed by the hybrid CTM, whereas PCA, PSD–AE, and JR produce more diffuse clusters. This illustrates that high MI and strong clustering are not always coupled and may reflect distinct inductive biases. In the Medium class the separation between representation families becomes sharper, with the hybrid CTM and Catch22 producing compact, well-isolated clusters, EEGNet yielding weaker separation, and PCA, PSD–AE, Hopf, and JR forming the lowest tier.

Taken together, the MI and clustering analyses reveal consistent stratification across methods. Data-driven and hybrid approaches align more closely with label structure, while constrained dynamical models tend to saturate and remain less separable. It is important to emphasise that the clustering scores are unsupervised validity indices. They capture the intrinsic tendency of latent spaces to form compact groups, which may align with class structure but do not directly measure class separability for the clinical tasks studied here.

5.3 CBMs yield independent features – hybrids show entanglement

Beyond raw information content, an important property of a latent representation is to provide disentangled or at least non-redundant features. Independent dimensions make interpretation easier, since each parameter can be mapped to a distinct physiological mechanism, while entanglement complicates this link by mixing multiple effects. To probe this aspect, we used an HSIC-based independence measure across methods.

In the *Small* class, nearly all methods exhibited very low pairwise dependence, suggesting that each latent captured a distinct aspect of the data (see Figure 4.3). This was expected for PCA, which produces orthogonal components, but it was equally true for CBMs. The notable exception was the hybrid `ctm_nn_avg`, which showed close to an order-of-magnitude higher dependence. While `ctm_nn`'s features carried the most task-relevant information, they were also the most entangled: changes in one inferred parameter tended to affect others. A plausible explanation is that the neural inference network did not learn to isolate the contribution of each physiological parameter, but instead adjusted groups of parameters jointly to match spectral patterns.

In the *Medium* class, where latent dimensionality is larger, we again found that most methods achieved strong feature independence (see Figure 4.11). Interestingly, the EEGNet autoencoder produced the lowest pairwise correlations of all, which seems surprising given its weak performance on other metrics. This pattern can be explained by its extreme underutilisation of latent units: only a small subset of dimensions was active, and these few carried largely non-overlapping information. The biophysical models (JR_pc and Hopf_pc) also produced highly independent features, in some cases even slightly outperforming PCA. By contrast, the PSD-AE and `catch22` showed moderate redundancy, reflecting overlapping spectral or statistical descriptors. Consistent with the Small class results, the hybrid `ctm_nn_pc` displayed the highest dependence among features.

Taken together, these results indicate that **CBMs do not inherently suffer a loss of feature independence compared with data-driven approaches**. On the contrary, their parameters often act as independent axes, much like PCA components, ensuring that each feature contributes unique information. This independence is advantageous for interpretability and downstream analysis. The caveat is that when we add a learned inference mechanism, as in `ctm_nn`, we sacrifice some of this modularity. Entanglement here appears to be the price for capturing more flexible, nonlinear relationships in the data.

Future work could investigate hybrid designs that explicitly encourage independence, for example through disentanglement regularisers, sparsity priors, or parameter orthogonalisation. Such strategies may help retain the mechanistic clarity of CBMs while still benefiting from the predictive gains of learned inference. More broadly, an open question is how much entanglement is acceptable for a model to remain interpretable in practice. Our findings suggest that hybrids strike a promising but still unresolved balance between modularity and task performance.

5.4 Geometric fidelity varies with model inductive bias

Why care about geometry? In this thesis we treat the PSD feature space as the original space whose neighborhood and distance relations we want to preserve. Faithful geometry matters because many downstream tools, graph constructions, manifold learning, and distance-based statistics, implicitly assume that the representation does not invent neighbors that were not present in the data manifold and does not pull apart points that should remain close.

A clear theme across weight classes is that geometric fidelity follows the objective. Methods that optimize variance capture or reconstruction align closely with the PSD manifold and therefore tend to preserve both local neighborhoods and global pairwise structure. This pattern is expected. Objectives based on explained variance or reconstruction penalize any twisting of the main spectral directions, so those axes stay intact. If the pipeline builds kNN graphs or summarizes pairwise distances, these methods give the most stable foundation.

The hybrid CTM occupies a middle ground. Adding amortised inference on top of a mechanistic manifold smooths the mapping from spectra to parameters and reduces obvious neighborhood breaks, yet it does not fully match the near-isometry of the variance/reconstruction baselines. This suggests a useful direction for model-based representations: learning helps mitigate geometric artifacts that arise from hard parameter fitting, but the learned mapping still inherits constraints from the biophysical coordinates it must predict.

Purely mechanistic CBMs show the widest spread and, in our setting, the weakest geometry. The likely driver is parameter degeneracy. When several parameter combinations explain similar spectra, geodesics in parameter space no longer reflect geodesics in PSD space. Per-channel parameterisations amplify this effect by introducing additional degrees of freedom that can trade off against one another. The implication is cautionary: distances between parameter vectors are not reliable proxies for spectral similarity unless identifiability is strong. If parameter-space analyses are the goal, stronger priors, identifiability constraints, or reduced parameterisations may be necessary.

EEGNet illustrates a different bias. It tends to keep local neighborhoods relatively intact while compressing global structure. This fits a design that concentrates information into a small active subset of units. Such embeddings can be serviceable for local retrieval or short-range smoothing, but long-range distances become less trustworthy. For tasks that depend on global geometry, they are a weaker fit.

Two caveats are important. First, geometry is not the same as informativeness. A representation can be almost isometric to PSD and still carry limited label information, and conversely a representation can distort distances while highlighting class-relevant contrasts. This explains why geometry and mutual information do not rank methods identically in our results. Second, PSD is a defensible and interpretable reference, but it is still a choice. Using averaged PSD for the Small group and stacked PSD for the Medium group matches the granularity of those settings, yet different reference features could shift the notion of what counts as a faithful embedding.

Taken together, the discussion points to a pragmatic guideline. If the analysis hinges on neighborhood graphs, distance statistics, or downstream algorithms that assume metric faithfulness, variance/reconstruction approaches are the safest default. When mechanistic interpretability is essential, the hybrid CTM provides a workable compromise that retains much of the PSD structure while keeping a physiological coordinate system. Pure CBMs remain valuable for hypothesis-driven analysis of specific mechanisms, but their parameter distances should not be used as stand-ins for spectral distances without additional safeguards. Future work should explore geometry-aware losses for CBMs, identifiability-strengthening priors, and joint objectives that trade off PSD faithfulness with task-relevant separation.

5.5 Representational similarity separates mechanistic and data-driven methods

Finally, we consider **how the different methods relate to one another**: do they learn essentially the same representation, or are they fundamentally capturing different "views" of the EEG data? By comparing pairwise alignment metrics, from linear CCA to nonlinear CKA, distance correlation, and Procrustes analysis, we uncovered a clear pattern of **representation families** (see Figures 4.6 and 4.14).

Methods that share similar objectives tend to produce strongly aligned latent spaces, whereas those with disparate inductive biases diverge significantly. For instance, the linear PCA and the PSD autoencoder are almost interchangeable in their representations. Adding the hybrid CTM (`ctm_nn`) to this mix, we find that `ctm_nn` clusters with the PCA/AE pair on many metrics, especially in the Medium (see Figure 4.14). This implies that **the CTM's learned parameter space, despite being constrained by a biophysical model, ends up encoding the data in a way that is largely congruent with the variance-maximising representations**.

By contrast, the dynamical CBMs occupy their own part of the representational landscape. In the Small class, Hopf and JR had relatively low alignment with the PCA/AE/hybrid trio, and importantly, they did not align strongly with each other either. The Wong–Wang model in Small was the most distinct of all. In the Medium class, the same broad divisions became even more pronounced. Under virtually every similarity metric, PCA, PSD-AE, and `ctm_nn_pc` form a tight cluster of mutually aligned representations, while JR_pc and Hopf_pc lie on the opposite end with minimal alignment to the rest.

The `catch22` features and EEGNet occupy an intermediate ground in Medium: they do not cluster with the PCA/hybrid group, but they are not as completely orthogonal as the mechanistic models. EEGNet, for instance, showed moderate linear CKA alignment with PCA, but differences appear in nonlinear metrics, reflecting its unique weighting of spectral features and its compression of global variance. `catch22` remained relatively weakly aligned with all others, which is an expected result given that its features are quite different descriptors of the signal than power spectra.

These representational (mis)alignments make a key point: **each method's inductive bias leads it to emphasize certain dimensions of variation in the EEG data at the expense of others, yielding distinct "views" of the same data**. Data-driven methods focus on axes of maximal variance or reconstruction fidelity, so they end up in agreement on the main spectral patterns. Mechanistic CBMs constrain representations to physiologically meaningful parameters, so they organize the data based on how those parameters can vary to produce the observed signals. An organisation that looks very different from a purely statistical decomposition.

Neither view is "wrong". They are simply different projections of the underlying multivariate brain activity. The divergence between representation families became sharper with the Medium models. When the EEG data were summarised only by an average PSD (Small class), even very different methods were constrained by the narrow scope of the input, and some commonalities emerged. But with full multi-channel spectra, there are far more ways to represent differences, and each method doubled down on its bias.

In practical terms, this means that **the choice of dimensionality reduction method will profoundly affect the structure of the features one obtains**. Our results highlight these contrasts but also some convergence: for example, the hybrid CTM demonstrates that a mechanistic model can be coaxed (via learning) into a representation that largely overlaps with a data-driven one. This is an encouraging sign for the idea of "hybrid methods", by combining mechanistic insights with flexible machine learning, we might get representations that are both interpretable and capture the bulk of statistically relevant structure.

Meanwhile, the clear isolation of the pure CBMs in representational similarity space underlines their unique value: they provide complementary perspectives that might reveal patterns invisible to PCA or neural nets. But that uniqueness comes with a trade-off: it can mean **diverging from**

the aspects of the data that are most useful for classification or other common analyses.

5.6 Summary - CBMs are efficient and interpretable but task-specific

In summary, our evaluation reveals a nuanced picture of CBMs as dimensionality reduction tools for EEG. In many fundamental respects, **CBM-based features behave comparably to those from conventional techniques**: they efficiently compress high-dimensional data, often preserve important geometric structure, and yield largely non-redundant representations. Crucially, they do so while providing interpretable parameters, something PCA or autoencoders mostly lack[35]. However, we also see where CBMs diverge from data-driven methods. Purely mechanistic models such as JR or Hopf often capture latent structures that are only weakly aligned with class-relevant patterns or with other standard reduction techniques. Their parameters remain interpretable but do not necessarily emphasize the same axes of variance that drive statistical discrimination in the data. This can be seen in their consistently lower mutual information and clustering scores, despite strong dimensional efficiency.

The hybrid approaches bridge this gap. The `ctm_nn` variants, for instance, not only maintained the efficiency and interpretability advantages of mechanistic models but also achieved high mutual information and strong clustering, aligning more closely with PCA and autoencoder families. This highlights that even modest learning components added to CBMs can markedly improve their practical utility as feature extractors, while still retaining a physiologically grounded latent space. However, the "Geometric and Feature Independence" metrics suggest caution. In many metrics the amortised CTM resembles data-driven methods and departs from its CMA-ES counterpart, which raises questions about its physiological grounding. A detailed analysis is outside the present scope and is left for future work.

Data-driven methods, by contrast, excelled in encoding task-relevant variance and in producing well-aligned latent geometries. PCA and PSD–AE consistently emerged as the closest pair across representational similarity analyses, reflecting their shared objective of variance preservation and reconstruction. EEGNet and Catch22, although distinctive in their organisation, showed more mixed performance: Catch22 benefited from handcrafted diversity, whereas EEGNet, despite its power, underutilised its latent capacity and yielded relatively poor information content.

Taken together, the results underscore that **the effectiveness of a dimensionality reduction method cannot be judged by any single criterion**. High efficiency does not guarantee class information, strong geometry preservation does not imply clustering, and high mutual information does not necessarily come with disentangled features. What emerges instead is a multidimensional profile: each method embodies a set of trade-offs dictated by its inductive bias.

For the purposes of clinical EEG, CBMs demonstrate that they can indeed function as dimensionality reduction techniques in the same analytical space as PCA or autoencoders. They compress the data effectively, offer largely non-redundant features, and sometimes preserve geometry exceptionally well. Where they lag is in encoding the strongest class-relevant signals, at least under the unsupervised training regime applied here. The hybrid approaches show the most promise in closing that gap.

It is also important to note that our evaluation of CBMs was carried out under a specific optimisation regime. Where appropriate, we relied on CMA-ES to fit model parameters, as it consistently balanced achieving a small reconstruction loss while being computationally tractable during testing. However, in the context of our experiments, the choice of optimizer is not a neutral detail: different inference strategies may shape the latent geometry in systematically different ways. For example, some methods might enforce smoother embeddings with stronger neighborhood preservation but at the cost of reduced cluster separability, while others could encourage sharper class-relevant boundaries at the expense of geometric fidelity. Our results therefore reflect one slice of a broader landscape: optimisation not only tunes models to data but also implicitly

biases which aspects of the latent space are emphasized, especially in the context of degenerate Computational Brain Models. This is something we already see when comparing the amortised and evolution strategy versions of our cortico-thalamic model. Exploring these interactions between model structure and optimisation criteria could open promising avenues for future work, revealing how different training dynamics sculpt the trade-offs we observed across efficiency, geometry, and informativeness.

Overall, this study positions computational brain models not merely as generative simulators but also as competitive alternatives to standard feature extraction tools. By reinterpreting them as dimensionality reduction techniques, we gain a new lens through which to evaluate their strengths and limitations, and to consider their integration with data-driven approaches. This discussion has highlighted the patterns, divergences, and complementarities across families of methods, providing a framework for thinking about how best to deploy CBMs for real-world EEG analysis.

In conclusion, these findings demonstrate that computational brain models **can indeed be understood and evaluated as dimensionality reduction methods for clinical EEG**. They achieve efficiency and interpretability on par with, and in some cases surpassing, conventional approaches, yet they diverge in how they capture class-relevant variance and organize latent space. This duality highlights both their promise and their limitations: while CBMs alone may not always optimize task performance, their unique inductive biases and physiologically grounded parameters offer complementary insights that data-driven methods cannot provide. Thus, the most compelling view is not CBMs versus data-driven techniques, but rather their integration as part of a broader representational toolkit for clinical neuroimaging.

Chapter 6

Conclusion and Future Work

This thesis set out to determine whether **biophysically guided computational brain models (CBMs)** can serve as effective dimensionality reduction methods for EEG data, and how they compare to conventional data-driven approaches. We evaluated interpretability, classification performance, robustness, and stability of various methods, through systematic experiments covering multiple representation scales (from compact 5-16 dimensional summaries up to medium-scale unconstrained latents). The results provide a nuanced understanding of the trade-offs involved, and they highlight both the promise and limitations of using mechanistic models for feature extraction.

Key Findings

Computational brain models (CBMs) are viable dimensionality-reduction frameworks for EEG that produce compact, interpretable latents. They use their capacity efficiently with very few inactive dimensions, whereas the EEGNet autoencoder concentrated information into a small active subset despite a larger bottleneck. Efficiency alone did not guarantee task utility: data-driven methods carried more label-relevant information and typically achieved slightly higher accuracy. PCA and PSD-AE best preserved the geometry of PSD space, and the amortised CTM recovered much of this structure while retaining mechanistic meaning. With increased capacity or per-channel parameters, CBM performance moved closer to data-driven baselines, which suggests the gap is architectural rather than fundamental. The hybrid ctm_nn ranked at or near the top on mutual information across both tasks and delivered competitive accuracy. Its representation aligned with the data-driven family while remaining somewhat more entangled than pure CBMs.

Trade-offs and Interpretation

The comparative analysis highlights a broader theme: **each category of model provides a different "view" of the data**, emphasising certain aspects while downplaying others. Data-driven methods excel at capturing dominant predictive patterns, but their features can be opaque. Pure CBMs provide interpretable mappings, with largely independent and physiologically plausible parameters, but may miss task-specific discriminative features. Hybrids and statistical baselines such as catch22 and PCA occupy a middle ground. PCA compresses spectra into low-dimensional coordinates that, while not biophysical, remain transparent and easy to analyse. Catch22 offers handcrafted time-series features that encode interpretable statistical properties of EEG without committing to a mechanistic model. Both therefore combine some of the clarity of structured features with the flexibility of data-driven extraction. Ultimately, no single method is universally best. The optimal choice depends on whether accuracy, interpretability, or balance is prioritized. Crucially, our work shows that these trade-offs are negotiable and through hybridisation, one can mitigate the weaknesses of each approach.

Limitations

This study has several limitations. Our evaluations were conducted on a specific EEG dataset (TUH) with two particular classification tasks, so the extent to which our conclusions generalize to

other datasets and tasks remains an open question. Second, the chosen CBMs are simplified representations of brain dynamics, unable to capture all EEG phenomena. This likely contributes to their lower classification performance in some settings. By contrast, non-CBM approaches such as PCA, autoencoders, or catch22 may derive part of their predictive power from non-neural sources present in EEG, including muscle artifacts or recording-specific structure. While useful for performance, this raises questions about physiological interpretability, and highlights an important trade-off between predictive accuracy and neural validity. Our hybrid model was a first step toward bridging this gap, but further work is needed to ensure its parameters remain disentangled and physiologically meaningful. Another limitation is that, while inputs ranged from PSDs to empirical channel data, all fittings and losses were defined spectrally, which may restrict the richness of information accessible to the models. Finally, computational cost is a practical limitation: optimisation-based inference for CBMs can be slow, and while amortised approaches alleviate this, scaling the same comparison to more sophisticated whole-brain models proved computationally intractable for us. Exploring such models would nevertheless be a highly valuable direction for future work. A further limitation is that our analyses focused exclusively on categorical classification. This choice makes conclusions dependent on the label definitions of the TUH dataset. Regression targets such as age estimation (objective) or standardised clinical measures (e.g., cognitive scores, disease severity) could provide more quantitative evaluation criteria and reduce dataset-specific bias.

Future Research Directions

Despite these challenges, the results suggest several promising avenues for future exploration:

- **Refining hybrid models:** More sophisticated integrations of CBMs with neural networks could yield physics-informed architectures that retain interpretability while enhancing flexibility.
- **Expanded validation:** Future studies should evaluate these methods on diverse datasets, tasks, and modalities (e.g. MEG, fNIRS), and assess robustness to domain shifts. Mechanistic priors may confer advantages in transferability, but this remains to be tested.
- **Improving mechanistic models:** Enriching CBMs with additional neural mechanisms, or combining multiple candidate models, could increase their scope. Maybe fine-tuning whole brain models for very compact latent space represenations could yield another promising way of increasing representation performance. This is not the only interesting way of improving mechanistic models: Simulation-based data augmentation is another promising avenue to strengthen hybrid training.

Closing Remarks

In summary, this thesis has shown that **computational brain models can be viewed and utilised as dimensionality reduction tools** for EEG, providing interpretable features that capture a substantial part of the signal's variance. While they do not always match the raw predictive power of unconstrained machine learning, their efficiency and interpretability are valuable, particularly in settings where physiological meaning is important. We also observed that design choices such as latent dimensionality and channel-specific parameterisation strongly affect performance, highlighting that CBMs are most effective when matched carefully to the task. Hybrid approaches, which combine mechanistic priors with data-driven inference, further indicate that the trade-off between interpretability and accuracy is not absolute, though this is only one direction among many. Overall, the results suggest that **CBMs are a viable addition to the toolbox of EEG feature extraction**, complementing data-driven methods by offering representations that are both compact and physiologically grounded.

Bibliography

[1] Lubba CH, Sethi SS, Knaute P, Schultz SR, Fulcher BD, Jones NS. catch22: CAnonical Time-series CHaracteristics. Data Mining and Knowledge Discovery. 2019 Nov;33(6):1821-52. Available from: https://doi.org/10.1007/s10618-019-00647-x.

[2] Assadzadeh S, Annen J, Sanz L, Barra A, Bonin E, Thibaut A, et al. Method for quantifying arousal and consciousness in healthy states and severe brain injury via EEG-based measures of corticothalamic physiology. Journal of Neuroscience Methods. 2023 Sep;398:109958.

[3] 10–20 system (EEG);. Accessed: 2025-09-07. https://en.wikipedia.org/wiki/10%E2%80%9320_system_(EEG).

[4] Glomb K, Cabral J, Cattani A, Mazzoni A, Raj A, Franceschiello B. Computational Models in Electroencephalography. Brain Topography. 2022;35(1):142-61. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8813814/.

[5] Alturki FA, AlSharabi K, Abdurraqeeb AM, Aljalal M. EEG Signal Analysis for Diagnosing Neurological Disorders Using Discrete Wavelet Transform and Intelligent Techniques. Sensors (Basel, Switzerland). 2020 Apr;20(9):2505. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7361958/.

[6] Zhu Y, Kandasamy R, Canham LJW, Western D. Window Stacking Meta-Models for Clinical EEG Classification. arXiv; 2024. ArXiv:2401.10283 [eess]. Available from: http://arxiv.org/abs/2401.10283.

[7] Madhavan S, Tripathy RK, Pachori RB. Time-Frequency Domain Deep Convolutional Neural Network for the Classification of Focal and Non-Focal EEG Signals. IEEE Sensors Journal. 2020 Mar;20(6):3078-86. Available from: https://ieeexplore.ieee.org/document/8913620.

[8] Mattioli F, Porcaro C, Baldassarre G. A 1D CNN for high accuracy classification and transfer learning in motor imagery EEG-based brain-computer interface. Journal of Neural Engineering. 2022 Jan;18(6).

[9] Jolliffe IT. Principal Component Analysis. Springer Series in Statistics. New York: Springer-Verlag; 2002. Available from: http://www.springer.com/statistics/statistical+theory+and+methods/book/978-0-387-95442-4.

[10] Lawhern VJ, Solon AJ, Waytowich NR, Gordon SM, Hung CP, Lance BJ. EEGNet: a compact convolutional neural network for EEG-based brain–computer interfaces. Journal of Neural Engineering. 2018 jul;15(5):056013. Available from: https://dx.doi.org/10.1088/1741-2552/aace8c.

[11] Jansen BH, Rit VG. Electroencephalogram and visual evoked potential generation in a mathematical model of coupled cortical columns. Biol Cybern. 1995 Sep;73(4):357–366. Available from: https://doi.org/10.1007/BF00199471.

[12] Robinson PA, Loxley PN, O'Connor SC, Rennie CJ. Modal analysis of corticothalamic dynamics, electroencephalographic spectra, and evoked potentials. Physical Review E. 2001 Mar;63(4):041909. Publisher: American Physical Society. Available from: https://link.aps.org/doi/10.1103/PhysRevE.63.041909.

[13] Wong KF, Wang XJ. A recurrent network mechanism of time integration in perceptual decisions. The Journal of Neuroscience: The Official Journal of the Society for Neuroscience. 2006 Jan;26(4):1314-28.

[14] Deco G, Kringelbach ML, Jirsa VK, Ritter P. The dynamics of resting fluctuations in the brain: metastability and its dynamical cortical core. Scientific Reports. 2017 Jun;7(1):3095. Publisher: Nature Publishing Group. Available from: https://www.nature.com/articles/s41598-017-03073-5.

[15] Kass RE, Amari SI, Arai K, Brown EN, Diekman CO, Diesmann M, et al. Computational Neuroscience: Mathematical and Statistical Perspectives. Annual Review of Statistics and Its Application. 2018 Mar;5(1):183-214. Available from: https://www.annualreviews.org/doi/10.1146/annurev-statistics-041715-033733.

[16] Montgomery RM. The Evolving Landscape of Computational Neuroscience: A Shift Towards Data-Driven and Integrative Approaches. Preprints; 2024. Available from: https://www.preprints.org/manuscript/202406.1203/v1.

[17] Obeid I, Picone J. The Temple University Hospital EEG Data Corpus. Frontiers in Neuroscience. 2016;Volume 10 - 2016. Available from: https://www.frontiersin.org/journals/neuroscience/articles/10.3389/fnins.2016.00196.

[18] de Diego SL. Automated Identification of Abnormal EEGs [Master's Thesis]. Philadelphia, PA: Temple University; 2017. Available from: https://www.isip.piconepress.com/publications/ms_theses/2017/abnormal/thesis/.

[19] Frässle S, Harrison SJ, Heinzle J, Clementz BA, Tamminga CA, Sweeney JA, et al. Regression dynamic causal modeling for resting-state fMRI. Human Brain Mapping. 2021 Feb;42(7):2159-80. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8046067/.

[20] Banville H, Chehab O, Hyvärinen A, Engemann DA, Gramfort A. Uncovering the structure of clinical EEG signals with self-supervised learning. arXiv; 2020. ArXiv:2007.16104 [stat]. Available from: http://arxiv.org/abs/2007.16104.

[21] Foumani NM, Mackellar G, Ghane S, Irtza S, Nguyen N, Salehi M. EEG2Rep: Enhancing Self-supervised EEG Representation Through Informative Masked Inputs. arXiv; 2024. ArXiv:2402.17772 [eess]. Available from: http://arxiv.org/abs/2402.17772.

[22] Puah JH, Guan C. EEGDM: EEG Representation Learning via Generative Diffusion Model; 2025. Available from: https://arxiv.org/html/2508.14086v3.

[23] David O, Friston KJ. A neural mass model for MEG/EEG: coupling and neuronal dynamics. NeuroImage. 2003 Nov;20(3):1743-55.

[24] Valdes-Sosa PA, Sanchez-Bornot JM, Sotero RC, Iturria-Medina Y, Aleman-Gomez Y, Bosch-Bayard J, et al. Model driven EEG/fMRI fusion of brain oscillations. Human Brain Mapping. 2009 Sep;30(9):2701-21.

[25] Sanz Leon P, Knock SA, Woodman MM, Domide L, Mersmann J, McIntosh AR, et al. The Virtual Brain: a simulator of primate brain network dynamics. Frontiers in Neuroinformatics. 2013 Jun;7. Publisher: Frontiers. Available from: https://www.frontiersin.org/journals/neuroinformatics/articles/10.3389/fninf.2013.00010/full.

[26] Moran RJ, Kiebel SJ, Stephan KE, Reilly RB, Daunizeau J, Friston KJ. A neural mass model of spectral responses in electrophysiology. NeuroImage. 2007;37(3):706-20. Available from: https://www.sciencedirect.com/science/article/pii/S1053811907004314.

[27] Newson JJ, Thiagarajan TC. EEG Frequency Bands in Psychiatric Disorders: A Review of Resting State Studies. Frontiers in Human Neuroscience. 2019 Jan;12. Publisher: Frontiers. Available from: https://www.frontiersin.org/journals/human-neuroscience/articles/10.3389/fnhum.2018.00521/full.

[28] Neural oscillation; 2025. Page Version ID: 1300108015. Available from: https://en.wikipedia.org/w/index.php?title=Neural_oscillation&oldid=1300108015.

[29] Engel AK, Fries P. Beta-band oscillations — signalling the status quo? Current Opinion in Neurobiology. 2010 Apr;20(2):156-65. Available from: https://www.sciencedirect.com/science/article/pii/S0959438810000395.

[30] Başar E. A review of alpha activity in integrative brain function: Fundamental physiology, sensory coding, cognition and pathology. International Journal of Psychophysiology. 2012 Oct;86(1):1-24. Available from: https://www.sciencedirect.com/science/article/pii/S0167876012003492.

[31] Welch P. The use of fast Fourier transform for the estimation of power spectra: A method based on time averaging over short, modified periodograms. IEEE Transactions on Audio and Electroacoustics. 1967 Jun;15(2):70-3. Available from: https://ieeexplore.ieee.org/document/1161901.

[32] Christ M, Braun N, Neuffer J, Kempa-Liehr AW. Time Series FeatuRe Extraction on basis of Scalable Hypothesis tests (tsfresh – A Python package). Neurocomputing. 2018 Sep;307:72-7. Available from: https://www.sciencedirect.com/science/article/pii/S0925231218304843.

[33] Gil Ávila C, Bott FS, Tiemann L, Hohn VD, May ES, Nickel MM, et al. DISCOVER-EEG: an open, fully automated EEG pipeline for biomarker discovery in clinical neuroscience. Scientific Data. 2023 Sep;10(1):613. Publisher: Nature Publishing Group. Available from: https://www.nature.com/articles/s41597-023-02525-0.

[34] Lagerlund TD, Sharbrough FW, Busacker NE. Use of principal component analysis in the frequency domain for mapping electroencephalographic activities: comparison with phase-encoded Fourier spectral analysis. Brain Topography. 2004;17(2):73-84.

[35] Principal components analysis (PCA) as a tool for identifying EEG frequency bands: I. Methodological considerations and preliminary findings | Request PDF. ResearchGate. 2025 Aug. Available from: https://www.researchgate.net/publication/298199832_Principal_components_analysis_PCA_as_a_tool_for_identifying_EEG_frequency_bands_I_Methodological_considerations_and_preliminary_findings.

[36] Hinton GE, Salakhutdinov RR. Reducing the Dimensionality of Data with Neural Networks. Science. 2006 Jul;313(5786):504-7. Publisher: American Association for the Advancement of Science. Available from: https://www.science.org/doi/10.1126/science.1127647.

[37] Kingma DP, Welling M. Auto-Encoding Variational Bayes. arXiv; 2022. ArXiv:1312.6114 [stat]. Available from: http://arxiv.org/abs/1312.6114.

[38] Jiang X, Zhao J, Du B, Yuan Z. Self-supervised Contrastive Learning for EEG-based Sleep Staging. arXiv; 2021. ArXiv:2109.07839 [cs]. Available from: http://arxiv.org/abs/2109.07839.

[39] Li W, Li H, Sun X, Kang H, An S, Wang G, et al. Self-supervised contrastive learning for EEG-based cross-subject motor imagery recognition. Journal of Neural Engineering. 2024 Apr;21(2).

[40] Weng W, Gu Y, Guo S, Ma Y, Yang Z, Liu Y, et al.. Self-supervised Learning for Electroencephalogram: A Systematic Survey. arXiv; 2024. ArXiv:2401.05446 [eess]. Available from: http://arxiv.org/abs/2401.05446.

[41] Breakspear M. Dynamic models of large-scale brain activity. Nature Neuroscience. 2017 Mar;20(3):340-52. Publisher: Nature Publishing Group. Available from: https://www.nature.com/articles/nn.4497.

[42] Robinson PA, Rennie CJ, Wright JJ. Propagation and stability of waves of electrical activity in the cerebral cortex. Physical Review E. 1997 Jul;56(1):826-40. Publisher: American Physical Society. Available from: https://link.aps.org/doi/10.1103/PhysRevE.56.826.

[43] Robinson PA, Rennie CJ, Rowe DL. Dynamics of large-scale brain activity in normal arousal states and epileptic seizures. Physical Review E. 2002 Apr;65(4):041924. Publisher: American Physical Society. Available from: https://link.aps.org/doi/10.1103/PhysRevE.65.041924.

[44] Wang XJ. Probabilistic Decision Making by Slow Reverberation in Cortical Circuits. Neuron. 2002 Dec;36(5):955-68. Publisher: Elsevier. Available from: https://www.cell.com/neuron/abstract/S0896-6273(02)01092-9.

[45] Deco G, Jirsa V, McIntosh AR, Sporns O, Kötter R. Key role of coupling, delay, and noise in resting brain fluctuations. Proceedings of the National Academy of Sciences. 2009 Jun;106(25):10302-7. Publisher: Proceedings of the National Academy of Sciences. Available from: https://www.pnas.org/doi/full/10.1073/pnas.0901831106.

[46] Breakspear M, Heitmann S, Daffertshofer A. Generative Models of Cortical Oscillations: Neurobiological Implications of the Kuramoto Model. Frontiers in Human Neuroscience. 2010 Nov;4. Publisher: Frontiers. Available from: https://www.frontiersin.org/journals/human-neuroscience/articles/10.3389/fnhum.2010.00190/full.

[47] Friston KJ, Harrison L, Penny W. Dynamic causal modelling. NeuroImage. 2003 Aug;19(4):1273-302.

[48] Zeidman P, Friston K, Parr T. A primer on Variational Laplace (VL). NeuroImage. 2023 Oct;279:120310. Available from: https://www.sciencedirect.com/science/article/pii/S1053811923004615.

[49] Hansen N. The CMA Evolution Strategy: A Tutorial. arXiv; 2023. ArXiv:1604.00772 [cs]. Available from: http://arxiv.org/abs/1604.00772.

[50] Wischnewski KJ, Eickhoff SB, Jirsa VK, Popovych OV. Towards an efficient validation of dynamical whole-brain models. Scientific Reports. 2022 Mar;12(1):4331. Publisher: Nature Publishing Group. Available from: https://www.nature.com/articles/s41598-022-07860-7.

[51] Momi D, Wang Z, Griffiths JD. TMS-evoked responses are driven by recurrent large-scale network dynamics. eLife. 2023 Apr;12:e83232. Publisher: eLife Sciences Publications, Ltd. Available from: https://doi.org/10.7554/eLife.83232.

[52] Moran RJ, Pinotsis DA, Friston KJ. Neural masses and fields in dynamic causal modeling. Frontiers in Computational Neuroscience. 2013 May;7. Publisher: Frontiers. Available from: https://www.frontiersin.org/journals/computational-neuroscience/articles/10.3389/fncom.2013.00057/full.

[53] Craik A, He Y, Contreras-Vidal JL. Deep learning for electroencephalogram (EEG) classification tasks: a review. Journal of Neural Engineering. 2019 Jun;16(3):031001.

[54] Sengupta B, Friston KJ, Penny WD. Efficient gradient computation for dynamical models. NeuroImage. 2014 Sep;98:521-7. Available from: https://www.sciencedirect.com/science/article/pii/S1053811914003097.

[55] Brodersen KH, Schofield TM, Leff AP, Ong CS, Lomakina EI, Buhmann JM, et al. Generative embedding for model-based classification of fMRI data. PLoS computational biology. 2011 Jun;7(6):e1002079.

[56] Pinotsis DA, Fitzgerald S, See C, Sementsova A, Widge AS. Toward biophysical markers of depression vulnerability. Frontiers in Psychiatry. 2022 Oct;13:938694. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9622949/.

[57] Raman S, Deserno L, Schlagenhauf F, Stephan KE. A hierarchical model for integrating unsupervised generative embedding and empirical Bayes. Journal of Neuroscience Methods. 2016 Aug;269:6-20.

[58] Dunstan DM, Richardson MP, Abela E, Akman OE, Goodfellow M. Global nonlinear approach for mapping parameters of neural mass models. PLOS Computational Biology. 2023 Mar;19(3):e1010985. Publisher: Public Library of Science. Available from: https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1010985.

[59] Ghosh S, Biswas D, Rohan NR, Vijayan S, Chakravarthy VS. Modeling of whole brain sleep electroencephalogram using deep oscillatory neural network. Frontiers in Neuroinformatics. 2025 May;19. Publisher: Frontiers. Available from: https://www.frontiersin.org/journals/neuroinformatics/articles/10.3389/fninf.2025.1513374/full.

[60] Sadegh-Zadeh SA, Sadeghzadeh N, Soleimani O, Shiry Ghidary S, Movahedi S, Mousavi SY. Comparative analysis of dimensionality reduction techniques for EEG-based emotional state classification. American Journal of Neurodegenerative Disease. 2024 Oct;13(4):23-33. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC11578865/.

[61] Anuragi A, Sisodia DS, Pachori RB. Mitigating the curse of dimensionality using feature projection techniques on electroencephalography datasets: an empirical review. Artificial Intelligence Review. 2024 Feb;57(3):75. Available from: https://doi.org/10.1007/s10462-024-10711-8.

[62] Schrimpf M, Kubilius J, Lee MJ, Ratan Murty NA, Ajemian R, DiCarlo JJ. Integrative Benchmarking to Advance Neurally Mechanistic Models of Human Intelligence. Neuron. 2020 Nov;108(3):413-23.

[63] Obeid I, Picone J. The Temple University Hospital EEG Data Corpus. Frontiers in Neuroscience. 2016;Volume 10 - 2016. Available from: https://www.frontiersin.org/journals/neuroscience/articles/10.3389/fnins.2016.00196.

[64] Gramfort A, Luessi M, Larson E, Engemann DA, Strohmeier D, Brodbeck C, et al. MEG and EEG data analysis with MNE-Python. Frontiers in Neuroscience. 2013;Volume 7 - 2013. Available from: https://www.frontiersin.org/journals/neuroscience/articles/10.3389/fnins.2013.00267.

[65] Klem GH, Lüders HO, Jasper HH, Elger C. The ten-twenty electrode system of the International Federation. The International Federation of Clinical Neurophysiology. Electroencephalography and Clinical Neurophysiology Supplement. 1999;52:3-6.

[66] Delorme A, Sejnowski T, Makeig S. Enhanced detection of artifacts in EEG data using higher-order statistics and independent component analysis. NeuroImage. 2007 Feb;34(4):1443-9. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2895624/.

[67] Jas M, Engemann DA, Bekhti Y, Raimondo F, Gramfort A. Autoreject: Automated artifact rejection for MEG and EEG data. arXiv; 2017. ArXiv:1612.08194 [stat]. Available from: http://arxiv.org/abs/1612.08194.

[68] Assadzadeh S, Annen J, Sanz L, Barra A, Bonin E, Thibaut A, et al. Method for quantifying arousal and consciousness in healthy states and severe brain injury via EEG-based measures of corticothalamic physiology. Journal of Neuroscience Methods. 2023 09;398:109958.

[69] Moran RJ, Stephan KE, Seidenbecher T, Pape HC, Dolan RJ, Friston KJ. Dynamic causal models of steady-state responses. NeuroImage. 2009 Feb;44(3):796-811. Available from: https://www.sciencedirect.com/science/article/pii/S1053811908010641.

[70] Deco G, Ponce-Alvarez A, Mantini D, Romani GL, Hagmann P, Corbetta M. Resting-state functional connectivity emerges from structurally and dynamically shaped slow linear fluctuations. The Journal of Neuroscience: The Official Journal of the Society for Neuroscience. 2013 Jul;33(27):11239-52.

Declaration

Use of generative AI I acknowledge the use of ChatGPT-5 (OpenAI, https://chatgpt.com) as a support tool for testing initial code implementations and understanding code bases, for proofreading, for suggesting relevant literature, and for exploring synonyms or alternative phrasings. I confirm that no content generated by AI has been presented as my own work.

Ethical Considerations All items in the provided ethics checklist were answered "No," indicating that the project does not involve any human or animal subjects, personal identifiable data, or other sensitive ethical issues. This research project uses the Temple University Hospital (TUH) EEG Abnormal corpus, which is a publicly available and fully anonymised set of EEG recordings [63]. No new data is collected from human participants, and all the data was originally gathered with Institutional Review Board approval and in accordance with the Declaration of Helsinki [63]. Because this dataset contains rigorously de-identified data in compliance with health data privacy standards (HIPAA), it satisfies privacy and data protection requirements [63]. As of this moment, there are no legal or intellectual property concerns: the dataset is open-access for research after signing a declaration of proper conduct, and the project employs only open-source tools, ensuring compliance with licensing and professional standards. In summary, the project raises no significant legal, social, ethical, or professional issues.

Sustainability All analyses in this project were conducted on existing university hardware using open-source software. Experiments were designed to minimise unnecessary re-runs, with efficient batch processing and caching to avoid repeated computations. We also employed modelling approaches that were computationally moderate, avoiding excessively resource-intensive methods. The overall environmental impact of the work is therefore negligible.

Availability of Data and Materials This study uses the Temple University Hospital (TUH) EEG Abnormal Corpus, which is publicly available for research under a data-use agreement [63]. All model code and scripts developed for this thesis are provided in the accompanying repository https://gitlab.doc.ic.ac.uk/lrh24/thesis.

Appendix A

Appendix

A.1 Computational Brain Model Fitting Results

This appendix presents detailed fitting results for all computational brain models tested on real EEG data. Each model was evaluated on 5 randomly selected EEG samples, with comprehensive comparison plots showing the model PSD fit quality, residuals, and frequency band analysis.

A.1.1 Model Overview

The following computational brain models were evaluated:

- **CTM-CMA-AVG**: Cortico-Thalamic Model with CMA-ES optimisation (channel-averaged)
- **CTM-NN-AVG**: Cortico-Thalamic Model with neural network inference (channel-averaged)
- **CTM-NN-PC**: Cortico-Thalamic Model with neural network inference (per-channel)
- **JR-AVG**: Jansen-Rit Model (channel-averaged)
- **JR-PC**: Jansen-Rit Model (per-channel)
- **Hopf-AVG**: Hopf Oscillator Model (channel-averaged)
- **Hopf-PC**: Hopf Oscillator Model (per-channel)
- **Wong-Wang-AVG**: Wong-Wang Model (channel-averaged)

A.1.2 Performance Summary

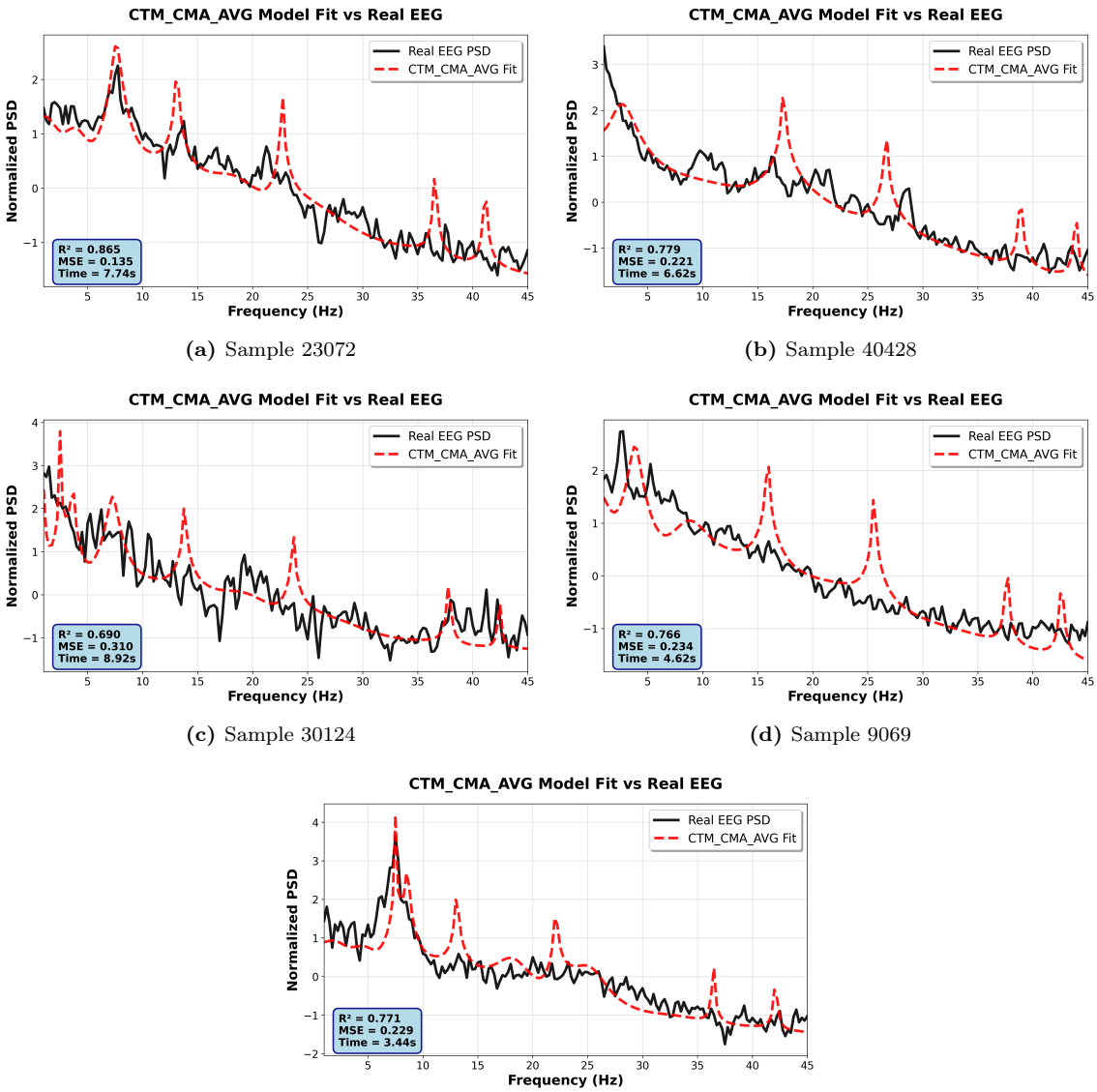
Table A.1: Computational Brain Model Performance Summary (5 samples)

Model	Avg Time (s)	Avg R^2	Avg MSE
CTM-CMA-AVG	6.27	0.774	0.226
CTM-NN-AVG	0.00	0.806	0.194
CTM-NN-PC	0.00	0.777	0.223
JR-AVG	2.33	0.920	0.080
JR-PC	47.17	0.910	0.090
Hopf-AVG	0.28	0.575	0.425
Hopf-PC	5.20	0.702	0.298
Wong–Wang-AVG	7.33	-0.946	1.946

Note. While most models produced stable and accurate spectral fits, the Wong–Wang implementation showed difficulties on the tested EEG samples. Negative R^2 values indicate that, for some recordings, fitted spectra deviated more from the data than a simple mean baseline, reflecting a mismatch between this model's dynamics and the spectral structure of the single-channel EEG samples.

A.1.3 Detailed Fitting Results

CTM-CMA-AVG Model Results



(a) Sample 23072

(b) Sample 40428

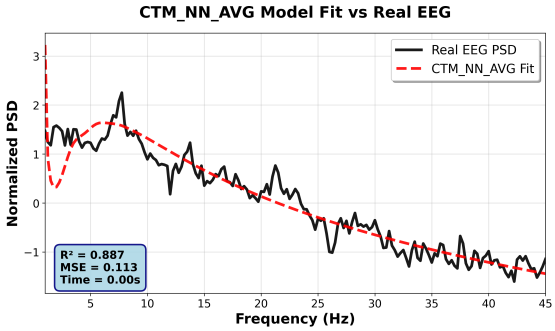
(c) Sample 30124

(d) Sample 9069

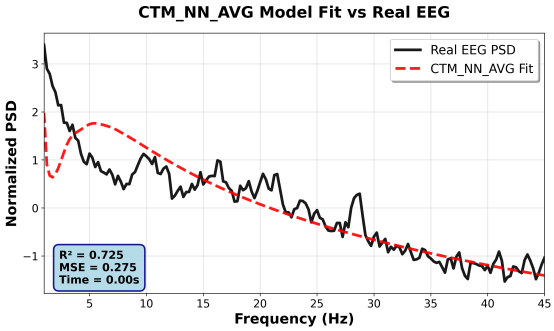
(e) Sample 7337

Figure A.1: CTM-CMA-AVG model fitting results for all test samples.

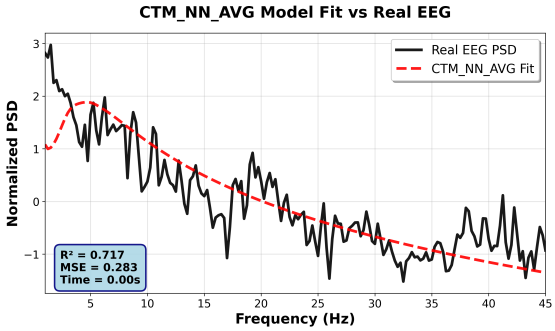
CTM-NN-AVG Model Results



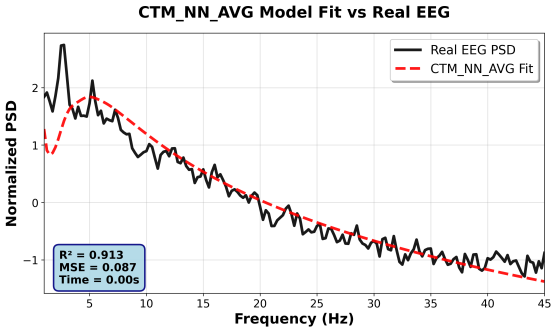
(a) Sample 23072



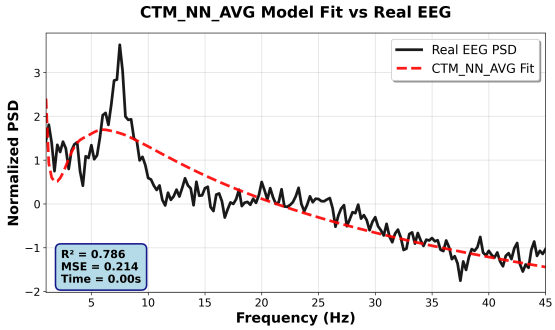
(b) Sample 40428



(c) Sample 30124



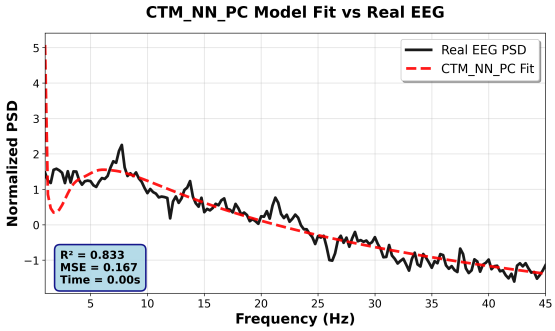
(d) Sample 9069



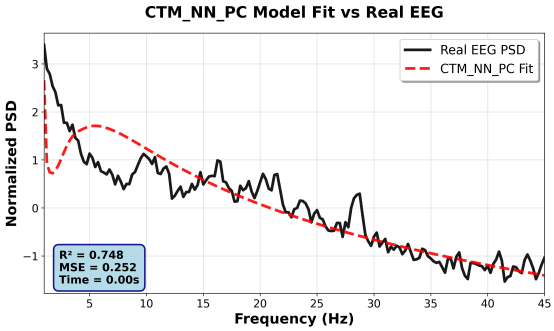
(e) Sample 7337

Figure A.2: CTM-NN-AVG model fitting results for all test samples.

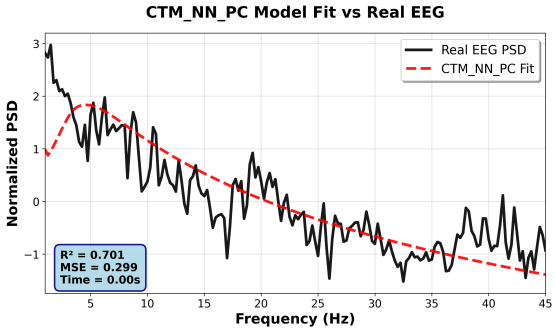
CTM-NN-PC Model Results



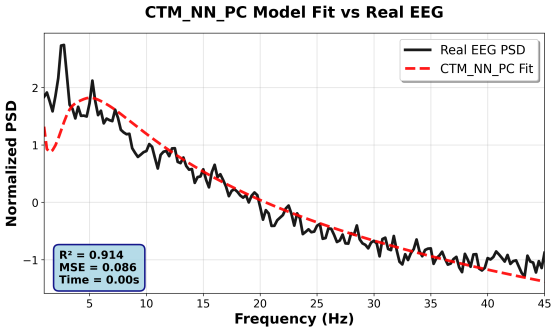
(a) Sample 23072



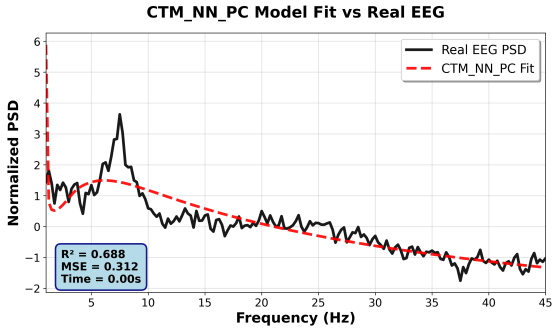
(b) Sample 40428



(c) Sample 30124



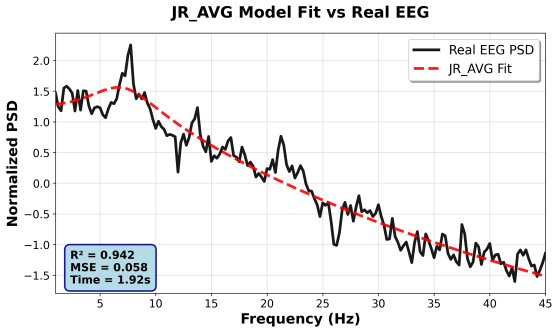
(d) Sample 9069



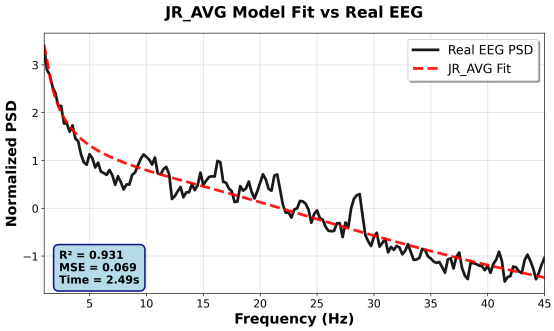
(e) Sample 7337

Figure A.3: CTM-NN-PC per-channel model fitting results. We visualize the fit by averaging the per-channel simulated and empirical PSDs.

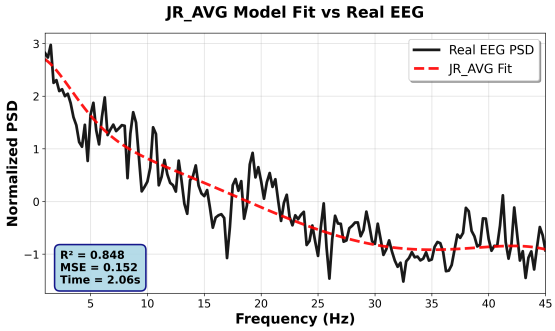
Jansen-Rit AVG Model Results



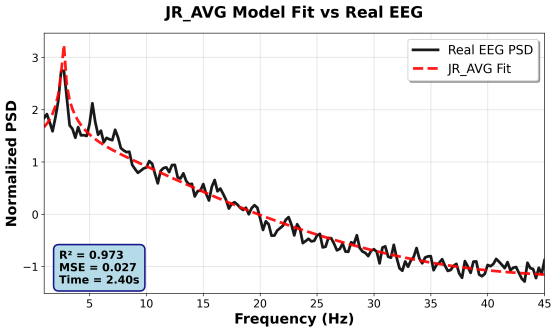
(a) Sample 23072



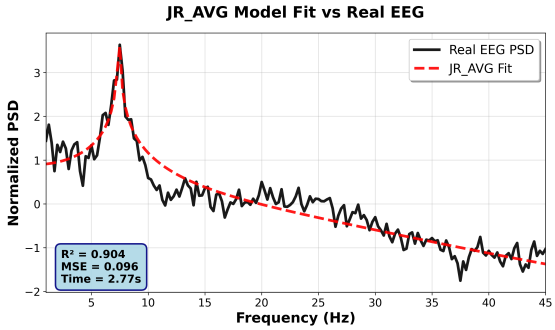
(b) Sample 40428



(c) Sample 30124



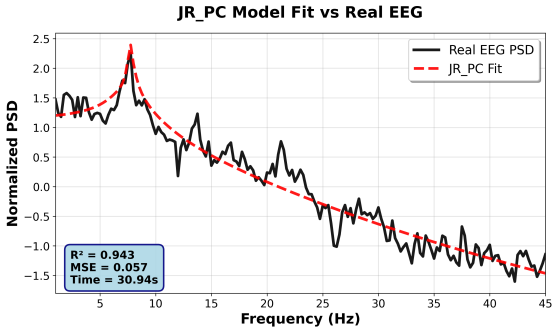
(d) Sample 9069



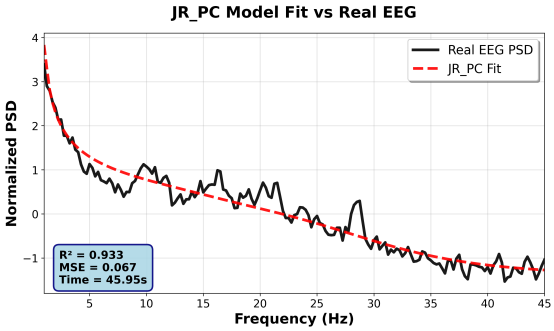
(e) Sample 7337

Figure A.4: Jansen-Rit channel-averaged model fitting results for all test samples.

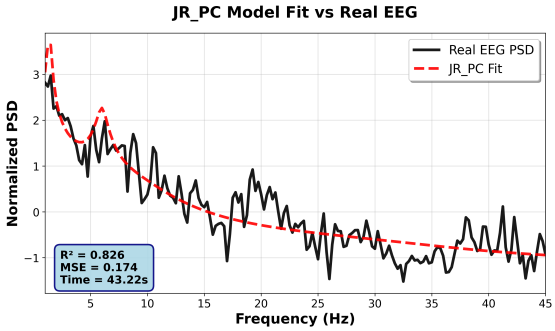
Jansen-Rit PC Model Results



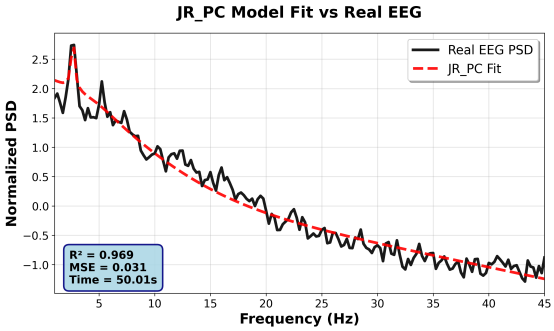
(a) Sample 23072



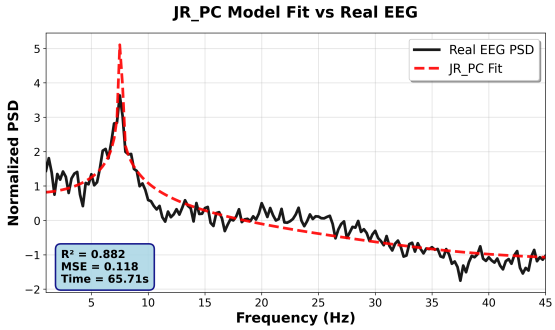
(b) Sample 40428



(c) Sample 30124



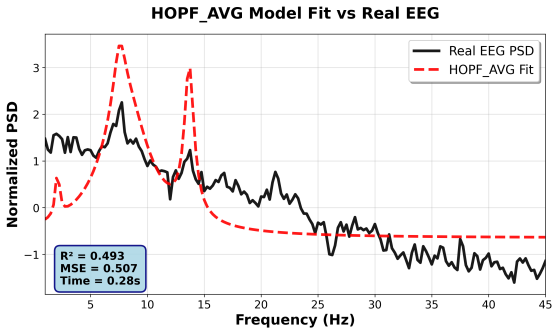
(d) Sample 9069



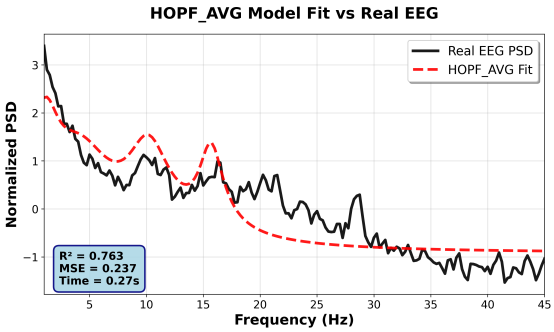
(e) Sample 7337

Figure A.5: Jansen-Rit per-channel model fitting results. Fits are shown per sample, aggregated for visualisation. We visualize the fit by averaging the per-channel simulated and empirical PSDs.

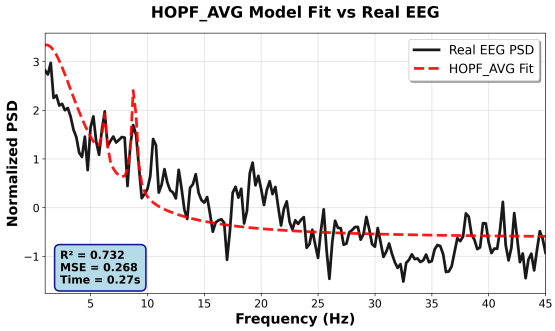
Hopf Oscillator Model Results



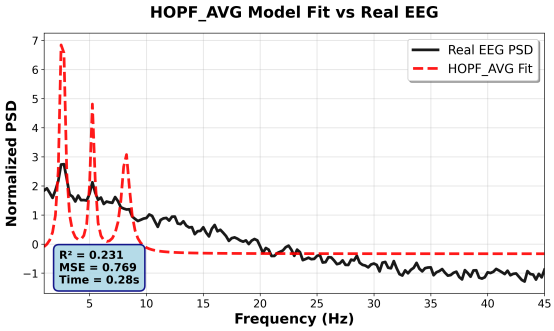
(a) Sample 23072



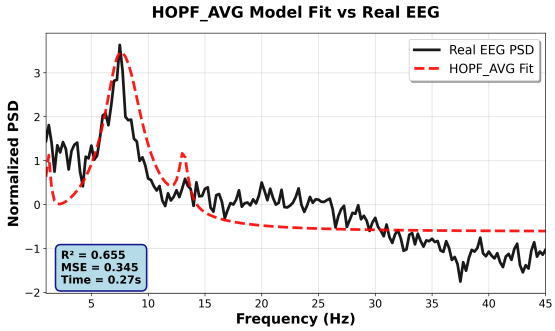
(b) Sample 40428



(c) Sample 30124



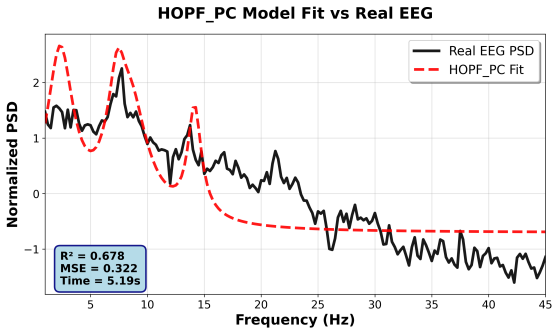
(d) Sample 9069



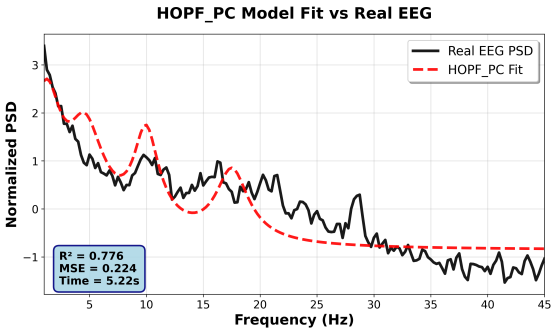
(e) Sample 7337

Figure A.6: Hopf oscillator channel-averaged model fitting results using Lorentzian spectral fitting approach.

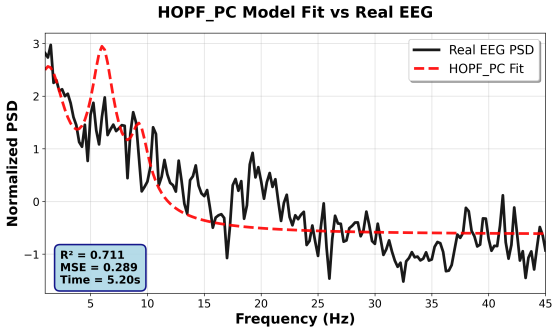
Hopf PC Model Results



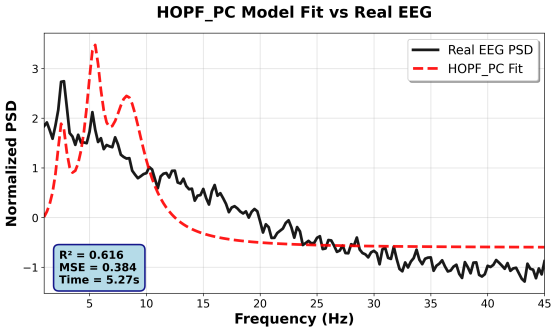
(a) Sample 23072



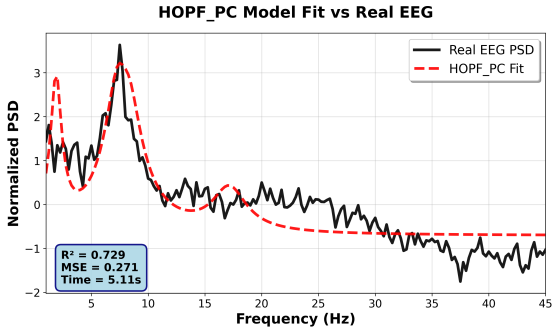
(b) Sample 40428



(c) Sample 30124



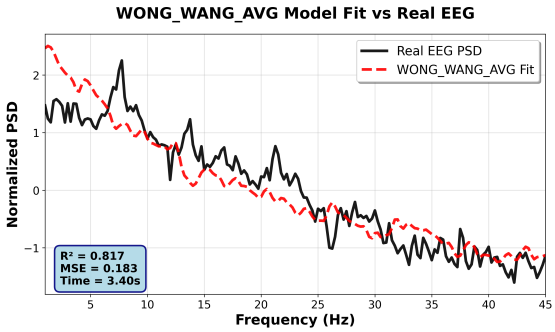
(d) Sample 9069



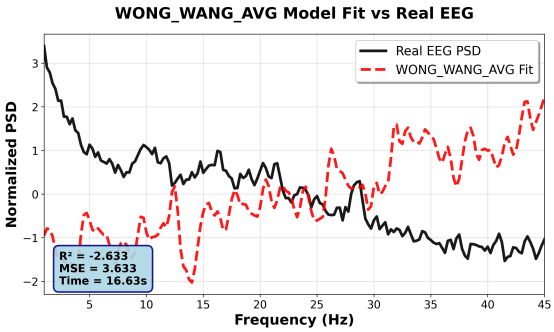
(e) Sample 7337

Figure A.7: Hopf oscillator per-channel model fitting results. Fits are shown per sample, aggregated for visualisation. We visualize the fit by averaging the per-channel simulated and empirical PSDs.

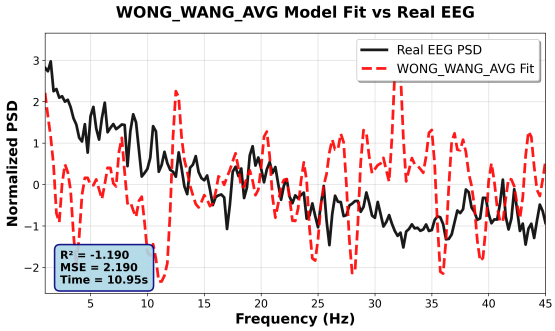
Wong-Wang Model Results



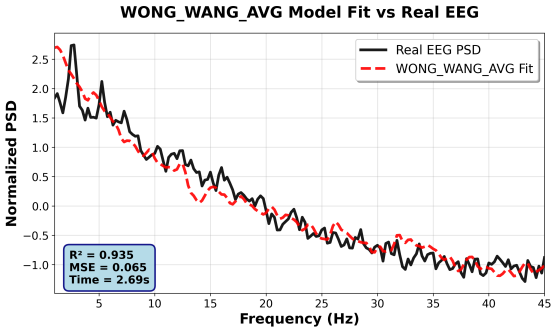
(a) Sample 23072



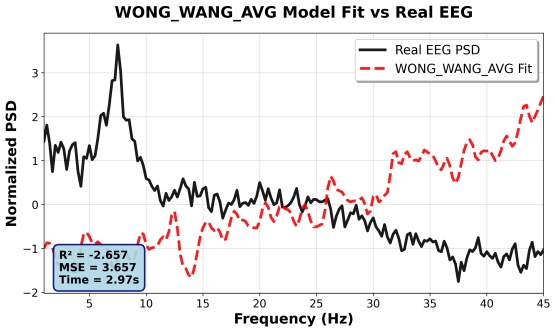
(b) Sample 40428



(c) Sample 30124



(d) Sample 9069



(e) Sample 7337

Figure A.8: Wong-Wang model fitting results using simulation-based PSD estimation.

A.2 Additional Latent Space Comparison Plots

Small Group

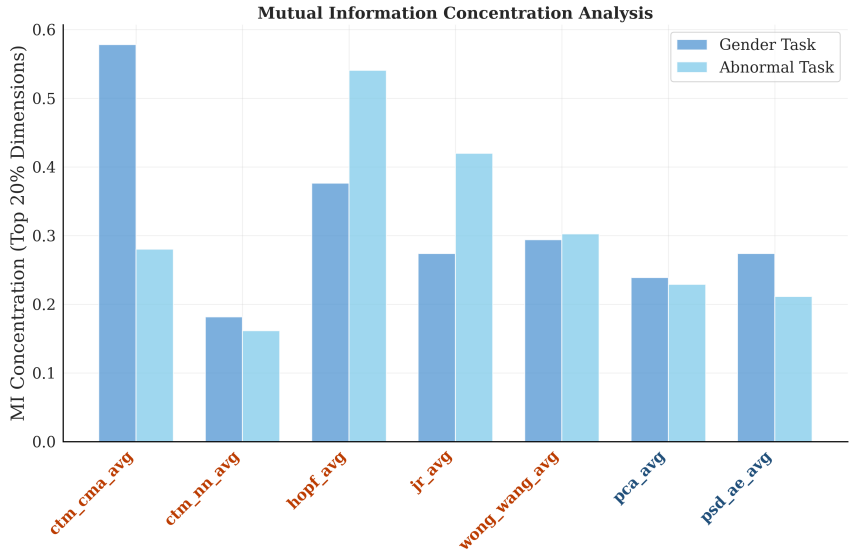


Figure A.9: Mutual Information in the top 20 percent of the Latent Dimension for the Small Group.

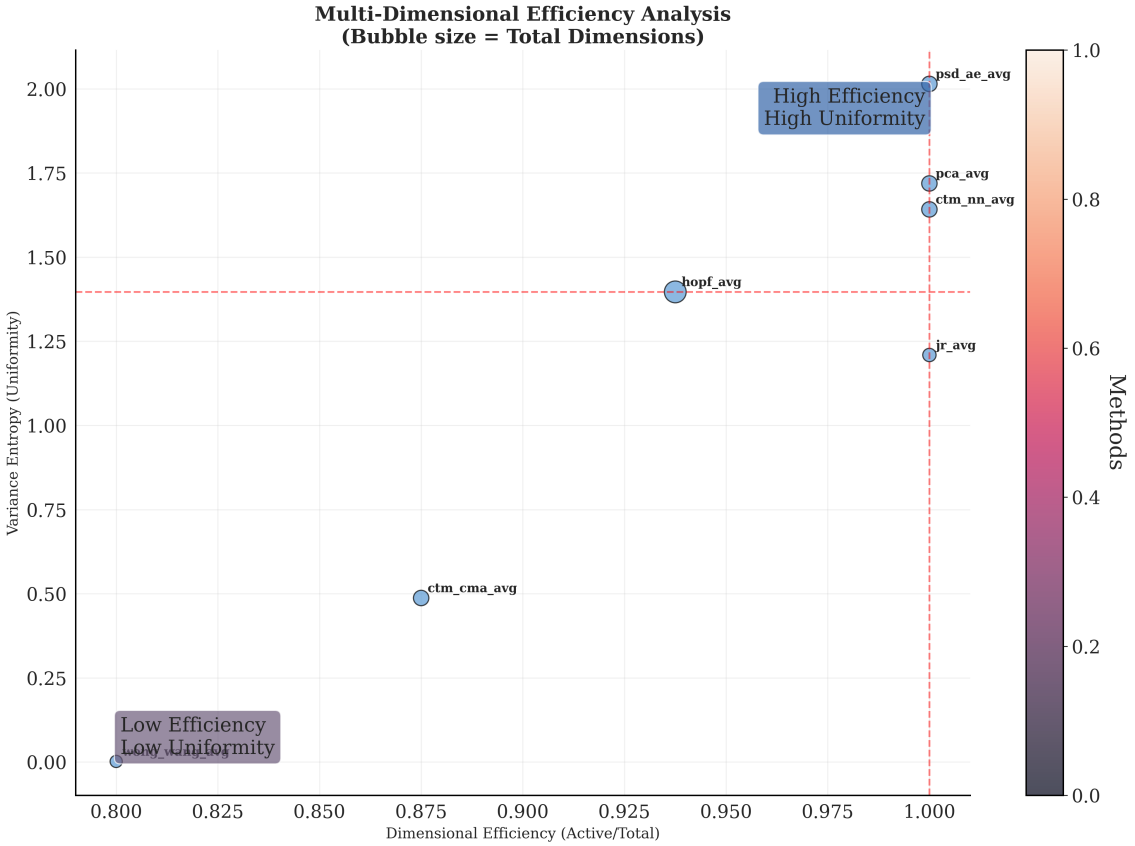


Figure A.10: Multi-Dimensional Efficiency Analysis for Small Group. Variance Distribution Entropy along dimensions vs. Dimensional Efficiency.

Medium Group

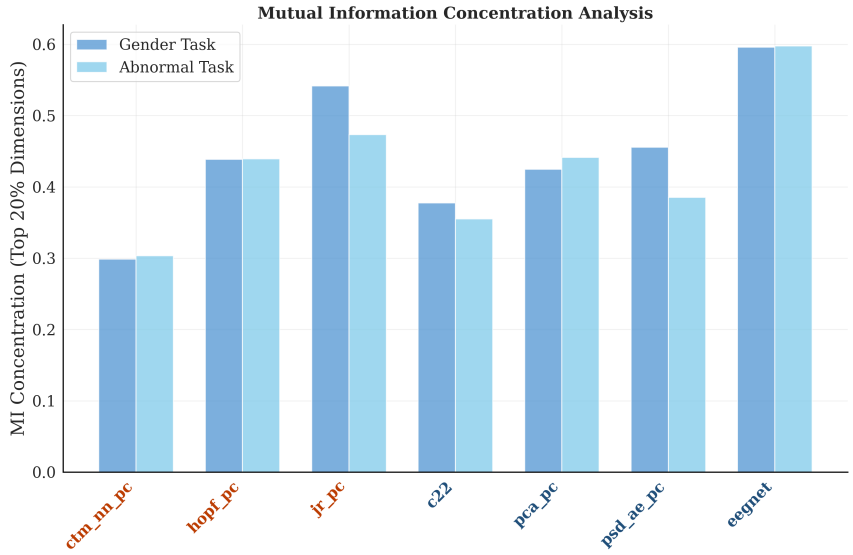


Figure A.11: Mutual Information in the top 20 percent of the Latent Dimension for the Medium Group.

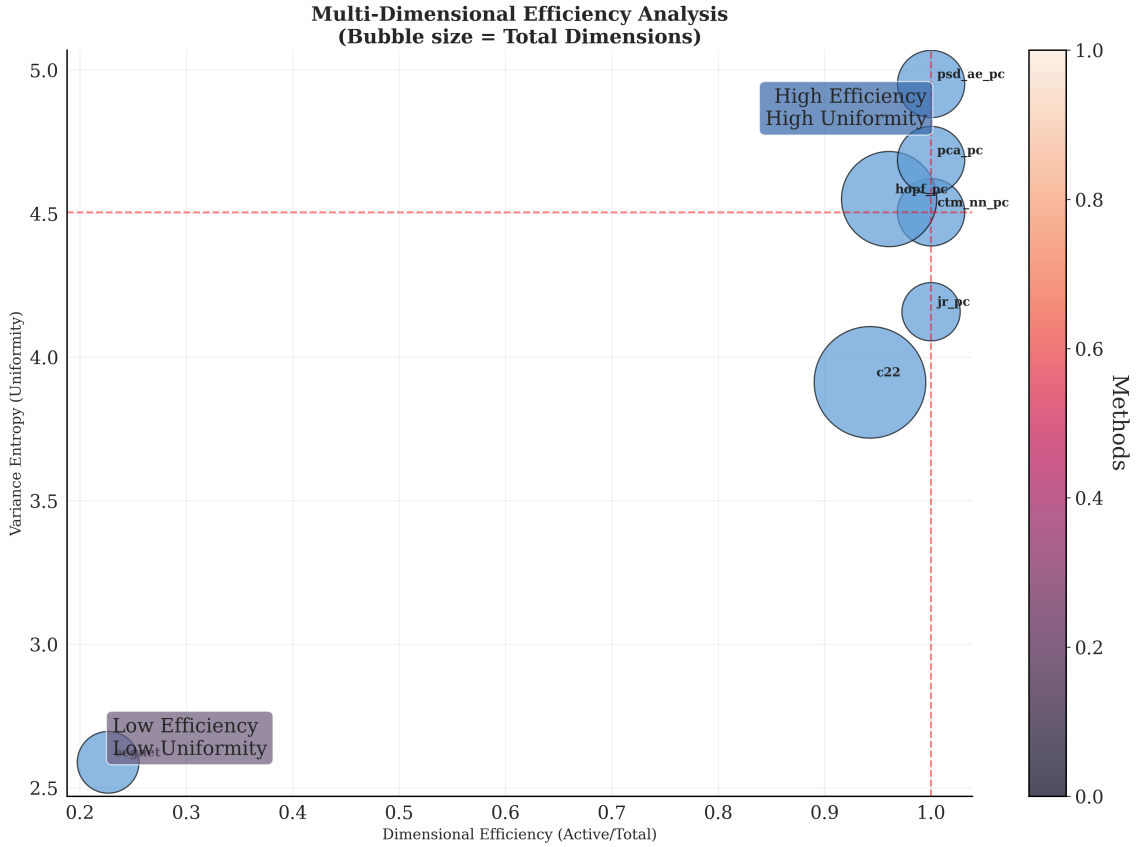


Figure A.12: Multi-Dimensional Efficiency Analysis for Medium Group. Variance Distribution Entropy along dimensions vs. Dimensional Efficiency.