SPATIAL ENCODING GEOMETRY DETERMINES THE MAPS AND TRANSITIONS RECOVERED BY DEEP EEG MICROSTATE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Unsupervised deep clustering is widely used for discovery in neuroscience, yet model choice is treated as an implementation detail. Models sharing the same data and evaluation metrics are assumed to recover equivalent structure. This assumption is tested in EEG microstate analysis, where discovered states feed biomarker studies. Two variational-deep-embedding models are compared on 203 paired eyes-closed recordings (LEMON). The comparison is centred on how the topography is spatially encoded: one model interpolates scalp topographies onto a 2-D image grid, the other learns on the anatomical electrode graph, with shared preprocessing, peak extraction, and evaluation throughout. As deployed, the two pipelines report a two-fold difference in temporal parameters: mean dwell 70.2 against 143.2 ms and occurrence 3.57 against 1.65 s^{-1}. Decoding each model's own recovered topographies with a common backfit removes that difference: mean dwell 70.2 against 71.0 ms and occurrence 3.57 against 3.37 s^{-1}. The reported divergence is therefore a property of the decode rather than of the encoding. What the encoding does change is the recovered maps. Under the common decode the graph model attains the higher backfit global explained variance, 0.618 against 0.542. Internal cluster metrics do not track this: the image model attains the higher silhouette (0.175 against 0.170, significant after Holm correction), and the indices are computed in different input geometries. These findings indicate that spatial encoding changes the recovered maps, the latent geometry, and the transition statistics of the sequences, but not the dwell and occurrence of the states, and that reported microstate parameters should state the backfit and the decode that produced them.

1 INTRODUCTION

EEG microstates are brief, quasi-stable configurations of the brain's electrical activity, characterised by a stable topography of the scalp potential field (Khanna et al., 2015). They typically last 60–120 ms (Michel & Koenig, 2018), and their sequence across the recording indexes the fast temporal organisation of spontaneous brain activity. The canonical analysis pipeline extracts snapshots at peaks of the global field power (GFP, the spatial standard deviation of the field) and clusters the resulting topographies into a small set of maps. The maps are then backfit to the continuous recording, and the global explained variance (GEV) is the standard measure of fit (Pascual-Marqui et al., 1995; von Wegner et al., 2018). Four canonical classes, labelled A through D, recur across recordings, cohorts, and laboratories, and constitute the de facto standard for resting-state analysis (Koenig et al., 2002).

Modified k-means and atomize-and-agglomerate hierarchical clustering, given the same GFP-peak topographies, return the same four canonical classes (Michel & Koenig, 2018). Their sequences are information-theoretically invariant, carrying statistically indistinguishable information-theoretic structure (von Wegner et al., 2018). For this reason, the choice of algorithm has carried little weight within the classical family. Variational deep embedding (VaDE) changes the terms by moving the clustering rule inside the generative model as a learnable Gaussian-mixture prior, and each topography is first mapped to a learned representation by an encoder-decoder pair (Jiang et al., 2017). This representation is a modelling choice rather than a property of the data. The same peak topog-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

raphy can be encoded as an interpolated two-dimensional image or as a signal on the anatomical electrode graph. The classical invariance result has nothing to say about this axis, because it varies the clustering rule while holding the input fixed. Whether the recovered temporal architecture is stable across such choices is, to our knowledge, untested at cohort scale. Convolutional variational autoencoders with Gaussian-mixture priors cluster the topographic maps and decode each cluster centre back into a scalp topography (Faremi & Longo, 2026). Systematic architecture search across this family, with multi-quadrant evaluation, maps how design choices trade internal cluster quality against fit and interpretability (Faremi et al., 2026). A related approach reduces topographies with a convolutional autoencoder and clusters the latent space in a separate step (Thukral et al., 2025). All of these models interpolate the topography onto a regular two-dimensional image grid before encoding. The interpolation invents pixels that no electrode recorded, and the grid locks the representation to one montage. However, these studies assess the quality of the maps, not of the dynamics they recover. Dwell times, occurrence rates, and transition structure remain unexamined.

These quantities are what the field reports, and they carry the comparisons made across disease, development, and treatment response (Khanna et al., 2015). They are sensitive measures. Microstate parameters differ between patients and controls in several neuropsychiatric conditions, and they change with age and with pharmacological intervention (Khanna et al., 2015; Koenig et al., 2002). A factor-of-two shift in dwell time or occurrence, introduced by the choice of model rather than by the brain, would pass through every downstream comparison that uses these parameters. The stakes attach to the choice of model, not only to the quality of the fit. Two analyses of the same recording can disagree by a factor of two before any scientific question is posed. Two variational deep embedding models are compared on 203 paired eyes-closed recordings from the LEMON dataset (Babayan et al., 2019). The models share the clustering prior, the training protocol, the preprocessing, the peak extraction, and every metric, and differ in how the topography is encoded; their published configurations differ additionally in latent dimension and loss terms. One model encodes the peak topography as an interpolated two-dimensional image. The other encodes the same peak topography as a signal on the anatomical electrode graph, with convolution over anatomically defined edges. Analyses that depend only on the shared inputs are identical across the two models, which shows that both pipelines consumed the same peaks, frames, and smoothing. Modified K-means, fit on the same peaks and backfit with the same smoothing, joins the comparison as the field's reference standard (Pascual-Marqui et al., 1995). The full protocol appears in Section 3.

Mean dwell is 70.2 ms for the graph model and 143.2 ms for the image model as their own pipelines report it, and occurrence is 3.57 s^{-1} against 1.65 s^{-1}. Every pair points the same way at every K in 3–20. That divergence does not survive a common decode. Re-decoding each model's own recovered topographies with the same smoother gives 70.2 against 71.0 ms in dwell and 3.57 against 3.37 s^{-1} in occurrence. The differences in the recovered maps and in the transition statistics do survive, and they do not point the same way. The image model attains the higher silhouette (0.175 against 0.170, Holm-adjusted $p = 0.015$) and the graph model the higher Calinski–Harabasz, yet the graph model wins backfit global explained variance for 100% of participants at every K under the common decode. Internal compactness does not explain that gap. Sequence complexity and surrogate-normalised recurrence statistics follow the reported pacing difference. The classical reference paces with both deep models once the decode is aligned. The Results section develops each point with effect sizes and confidence intervals. The research question is whether two deep generative embedding models trained on the same microstate data, differing in how the topography is spatially encoded, recover the same structure. The research hypothesis is that the spatial encoding changes the recovered maps, the latent geometry, and the transition statistics of the sequences, while a common decode brings the dwell and occurrence of the states together, so a comparison of those parameters across pipelines is meaningful only under a common decode. The classical invariance result serves as the null for the temporal measures. The hypothesis follows from the argument above. Representation is a modelling choice, and that choice should leave a measurable trace in what the models recover. The study makes four contributions.

1. A cohort-scale paired quantification of what spatial encoding changes in deep microstate models, across 203 recordings and every K in 3–20.

2. A decode-controlled design that separates pipeline configuration from spatial encoding, with input sameness proven inside the data rather than assumed.

3. A demonstration that the reported temporal divergence between two deep microstate pipelines is a property of the decode, not of the encoding.
4. A validity checklist for the temporal dynamics of deep microstate models.

Section 2 reviews related work. Section 3 specifies the controlled protocol and the formal background. Section 4 reports the results. Section 5 discusses implications, limitations, and conclusions.

2 Related Work

Microstate estimation was formalised as the segmentation of the scalp field into quasi-stable states (Lehmann et al., 1987; Pascual-Marqui et al., 1995) and has converged on a small family of clustering algorithms operating on GFP-peak topographies, with modified K-means and atomize-and-agglomerate hierarchical clustering as the reference methods (Murray et al., 2008; Khanna et al., 2015; Michel & Koenig, 2018). Four canonical classes recur across cohorts (Koenig et al., 2002), and global explained variance is the standard measure of fit (Murray et al., 2008). Because the choice among the classical algorithms leaves the information-theoretic structure of the label sequences unchanged (von Wegner et al., 2018), the choice of algorithm has carried little weight within this family, although PCA- and ICA-based variants do alter the dynamics (von Wegner et al., 2018).

Variational deep embedding couples a variational autoencoder to a Gaussian-mixture prior so that clustering and representation learning are trained jointly (Jiang et al., 2017). Along this line, the variational autoencoder has been coupled to a Gaussian-mixture prior for joint manifold learning and clustering of EEG topographic maps (Faremi & Longo, 2026), and a systematic architecture search has mapped how design choices trade the quadrants of cluster validity against reconstruction quality and interpretability (Faremi et al., 2026). A related approach reduces topographies with a convolutional autoencoder before clustering the latent space (Thukral et al., 2025). Recurrent models have explored the temporal dynamics of microstates (Sikka et al., 2020).

Graph neural networks generalise convolution to irregular domains and have been applied to EEG with electrodes as graph nodes (Klepl et al., 2024). The operators used in this study are the graph convolutional network (Kipf & Welling, 2017), graph attention (Veličković et al., 2018), and Chebyshev spectral filtering (Defferrard et al., 2016). To our knowledge, microstate discovery has not previously represented the montage as a graph. Graph embeddings of EEG connectivity have been used for representation learning and for comparative analyses of structural and functional networks (Almohammadi & Wang, 2024). The temporal structure of microstate sequences has been characterised with information-theoretic measures that are invariant across classical clustering algorithms (von Wegner et al., 2018; 2023). Dynamical analyses have refined the surrogate methodology, with first-order Markov surrogates as the appropriate null (Jajcay & Hlinka, 2023; von Wegner et al., 2025). The resulting transient structure relates to the fast transient networks of spontaneous brain activity (Baker et al., 2014). Dwell-time structure has been modelled explicitly with hidden semi-Markov models (Borst & Anderson, 2015), and group-level microstate analyses carry documented biases that individual-level designs avoid (Murphy et al., 2023). The current state of the methodology is reviewed by Michel et al. (2024). What remains open is whether deep generative models recover dynamics that agree with each other and with the classical reference. This study occupies that open position. The comparison holds the modelling framework and the evaluation fixed, varies the spatial representation of the topography, and reports the two published pipeline configurations as a second axis. The classical reference enters the design rather than the discussion. Temporal validity is reported alongside internal cluster quality, which the deep microstate literature has not done at cohort scale.

3 A Controlled Protocol for Comparing Deep Microstate Models

3.1 Formalism

At each time point t the scalp field is a vector $\mathbf{x}_t \in \mathbb{R}^C$ over C channels. The global field power at t is the spatial standard deviation σ_t of $\mathbf{x}_t$. The analysis operates on the topographies at local maxima of σ_t. A microstate map is a unit-norm template $\mathbf{m}_k$. The backfit assigns each time point to the map

Table 1: The two models share the clustering prior, the training protocol, and the evaluation. They differ in the spatial encoding of the topography and, additionally, in latent dimension and loss terms.

Component	CNN-based model (Conv-VaDE)	Graph-based model (Graph-VaDE)
Input sample	40×40 interpolated topomap	61-dim vector on electrode graph
Spatial prior	azimuthal equidistant grid	$A_{ij} = \exp(-d_{ij}^2/\delta^2)$, thresholded
Encoder	Conv2d stack + BN	graph conv (GCN/GAT/ChebNet; depth 2 reported)
Decoder	transposed-convolution stack	shared node-wise MLP
Clustering prior	learnable GMM (VaDE), closed-form 3-term KL	
Polarity invariance	flip augmentation + polarity-abs reconstruction + latent regulariser	
Training	VAE pretrain → K-means init → joint VaDE; deterministic seeds	

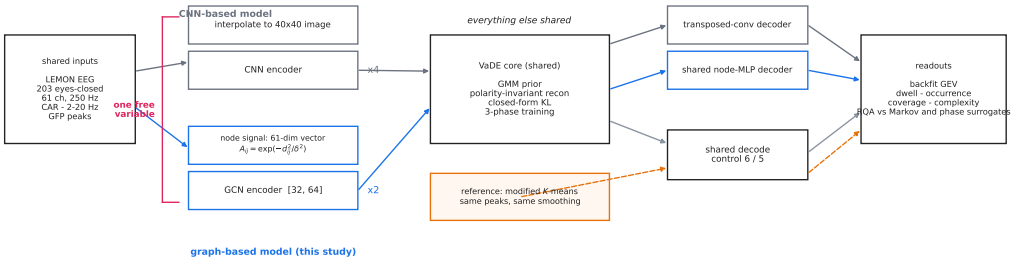

Figure 1: The controlled comparison. Every stage outside the two encoding branches is shared (grey): the preprocessing, the peak extraction, the Gaussian-mixture prior, the three-phase training, and every metric. The decode control applies one smoothing to both models' recovered topographies. Modified K-means (orange) joins on the same peaks and the same smoothing as the classical reference.

of maximum absolute spatial correlation. The global explained variance of a set of maps is

$$\mathrm{GEV} = \frac{\sum_t \sigma_t^2 \, |\mathrm{corr}(\mathbf{x}_t, \mathbf{m}_{l(t)})|^2}{\sum_t \sigma_t^2}, \tag{1}$$

where $l(t)$ is the map assigned at time t. Polarity invariance enters through the absolute correlation. A topography and its sign flip receive the same label.

VaDE extends deep embedded clustering (Xie et al., 2016) with a variational generative model (Kingma & Welling, 2014; Jiang et al., 2017). Each snapshot $\mathbf{x}$ is generated from a Gaussian-mixture prior over a latent variable $\mathbf{z}$. The generative model is $p(\mathbf{x}, \mathbf{z}, c) = p(c)\, p(\mathbf{z} \mid c)\, p(\mathbf{x} \mid \mathbf{z})$. The component c selects a mixture element. A variational autoencoder parameterises the posterior $q(\mathbf{z} \mid \mathbf{x})$. Training maximises the evidence lower bound, $\mathcal{L} = \mathbb{E}_q[\log p(\mathbf{x} \mid \mathbf{z})] - \mathrm{KL}\big(q(\mathbf{z}, c \mid \mathbf{x}) \,\|\, p(\mathbf{z}, c)\big)$. The mixture prior and the component assignment are learned jointly with the encoder and the decoder (Jiang et al., 2017). The two models of this study instantiate the same objective. They differ in the operator that encodes $\mathbf{x}$ and produces the reconstruction, and additionally in latent dimension and loss terms, as specified in Section 3.

3.2 Design

The study compares two deep generative models based on the variational deep embedding (VaDE) architecture for EEG microstate discovery. The comparison follows a controlled-change design centred on one variable: how the topography is spatially encoded. The two models follow their published configurations, which differ additionally in latent dimension (10 versus 32), in loss terms, and in their backfit and decode settings. Preprocessing, GFP-peak extraction, and every metric are shared. Both models are trained with the same protocol, so the comparison is between two spatial representations rather than between modelling frameworks. The shared stages are preprocessing

(common average reference, 2–20 Hz zero-phase FIR filtering), GFP-peak extraction, and every metric. The graph-based model, the classical reference, and the decode control all use the same smoothing of the backfit labels (half-window 6, minimum segment 5 samples). The classical reference, modified K-means, is fit on the same peaks and backfit with the same smoothing. Table 1 lists the components, and Figure 1 summarises the design.

3.3 DATA

The study made use of 203 eyes-closed resting-state EEG recordings from the LEMON dataset (Babayan et al., 2019). The publicly distributed preprocessed archive contains exactly 203 participants, each with one eyes-closed recording. The preprocessed recordings are sampled at 250 Hz with 61 EEG channels, except for two that the archive serves at 100 Hz. For those two recordings the backfit smoothing and the recurrence window are specified in samples and therefore span a coarser physical scale than elsewhere, while the recurrence embedding lag is rate-adapted and holds at 40 ms at both rates. They are retained, and excluding them changes no reported statistic to the precision given here: the median dwell moves from 67.4 to 67.3 ms and the in-band share from 94.1 to 95.0% under the graph-based model, whose dwell range is 59 to 169 ms over the full cohort and 59 to 97 ms without them. The cohort coincides with that of (Faremi & Longo, 2026). The dataset is public and anonymised, and the original collection was conducted with ethical oversight and informed consent (Babayan et al., 2019).

3.4 MODELS AND TRAINING

The CNN-based model interpolates the topography onto a 40-by-40 grid and encodes it with a stack of convolutional layers (Faremi & Longo, 2026). A transposed-convolution decoder maps the latent state back to the grid. A Gaussian-mixture prior sits on the latent space. The graph-based model encodes the 61-dimensional peak vector as a signal on the anatomical electrode graph. Nodes sit at sensor positions. Edges are weighted by $A_{ij} = \exp(-d_{ij}^2/\delta^2)$. The scale δ is the standard deviation of pairwise distances. An edge is retained when $d_{ij} \leq 0.4 \max d$. The encoder applies graph convolution layers to this signal. The configuration follows the architecture search reported in the companion study, where an Optuna search with a TPE sampler and Hyperband pruning (Akiba et al., 2019) was run on a ten-participant LEMON cohort. It is the GCN variant with two layers of width 32 and 64, latent dimension 10, flatten readout, and dropout 0.15. The same fixed configuration is used for every participant and at every K from 3 to 20; no per-participant search was run. A shared node-wise multilayer perceptron decodes each node signal to a channel voltage. Both models share the VaDE objective. The objective is a polarity-invariant reconstruction term plus a closed-form Gaussian-mixture KL divergence. Training proceeds in three phases. The first pretrains the variational model. The second initialises the mixture with K-means. The third optimises the full objective jointly. All runs use deterministic seeds. A held-out fraction of 0.1 of the peaks is drawn at random and used for early stopping and for the validation metric. Polarity invariance is enforced in both models by three mechanisms. The training set is augmented with sign-flipped topographies. The reconstruction loss is polarity-invariant. An encoder regulariser penalises the latent deviation between a topography and its sign flip (Faremi & Longo, 2026). The CNN-based model follows the published pipeline (Faremi & Longo, 2026). Its configuration incorporates the interpolation onto the image grid as published. The training configuration is given in Appendix B.

3.5 METRICS AND STATISTICS

The main measure evaluated is the backfit global explained variance at $K = 4$. It is computed on the continuous recording, with each sample assigned to the template of maximum absolute spatial correlation and each template's share weighted by the squared global field power. Every sample is assigned at $K = 4$, so no unlabelled fraction enters the measure. The definition was cross-checked across the two codebases during the audit of the comparison. The posterior GEV of the model is used for model selection only; the backfit GEV is used for every cross-model claim. Secondary measures are the temporal parameters of the smoothed backfit. These are mean dwell time, occurrence rate, and coverage. Internal cluster validity is reported alongside. These are silhouette, Calinski–Harabasz, and Davies–Bouldin. Each index is computed in its own model's input geometry, the 1600-pixel interpolated grid for the CNN-based model and the 61-dimensional channel

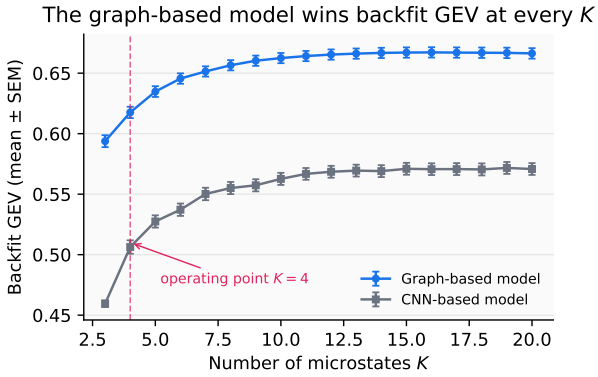

Figure 2: Backfit GEV across K. The graph-based model (blue) attains the higher mean at every K and wins all 203 paired comparisons; the dashed line marks the operating point $K = 4$.

vector for the graph-based model, so the magnitudes are not comparable across the two models and only their directions are read. Sequence complexity is quantified by Shannon entropy, entropy rate, and Lempel–Ziv complexity. Recurrence structure is quantified by determinism and by a z-score against 99 surrogates (von Wegner et al., 2025). Recurrence is computed on five representations of the sequence. These are the symbolic sequence, the jump sequence, the continuous GFP series, the multichannel series, and the latent trajectory. The symbolic and jump sequences are tested against first-order Markov surrogates, which preserve the transition structure of the sequence. The continuous, multichannel, and latent representations are tested against phase-randomised Fourier surrogates, with one shared phase vector across channels and latent dimensions so that the cross-spectrum is preserved. The Markov order of the label sequence is tested with a likelihood-ratio statistic on the full sample-level sequence, with self-transitions retained and degrees of freedom $(K-1)^2$ and $K(K-1)^2$. Long-range dependence is quantified by detrended fluctuation analysis of the per-state occupancy. The analysis is implemented in Python with MNE-Python (Gramfort et al., 2013) and PyTorch.

All comparisons are paired by participant. The main statistical test is the Wilcoxon signed-rank test. P-values are Holm-corrected across the 17 tested paired measures at $K = 4$. Effect sizes are rank-biserial correlations with bootstrap confidence intervals. The bootstrap uses 10,000 resamples over participants. Discordant-pair counts are reported next to every effect. Curves over K report the mean and the standard error of the mean. The per-measure sample size is stated wherever it falls below 203. Internal-validity metrics for both models derive from the per-participant training records and cover all 203 participants. Analyses that depend only on the shared inputs must agree across the two models if the pipelines consumed the same recordings and peaks. Two such analyses are the recurrence measures on the continuous GFP series and on the multichannel series. These measures are identical across the two models at $K = 4$ for all 203 participants. The pipelines therefore consumed the same inputs. This check covers the shared inputs; the label-level decode is addressed by the control in Section 4. It is reported before any model-dependent result.

4 Results

The question under examination is whether two models that share their inputs and their evaluation recover the same structure. The results separate two answers. The temporal divergence their pipelines report is a property of the decode, and a common decode removes it. The advantage in explained variance survives that control. Figure 2 shows the backfit global explained variance for both models at every K from 3 to 20. The graph-based model attains the higher mean at every K. At $K = 4$ the means are 0.618 and 0.506. The graph-based model wins all 203 paired comparisons at every K. Discordant pairs number zero throughout.

Figure 3 shows the per-participant temporal parameters at $K = 4$ as the two pipelines report them. Mean dwell is 70.2 ms for the graph-based model and 143.2 ms for the CNN-based model. The difference is 72.96 ms with a bootstrap 95% confidence interval of [70.6, 75.6] ms. Occurrence is

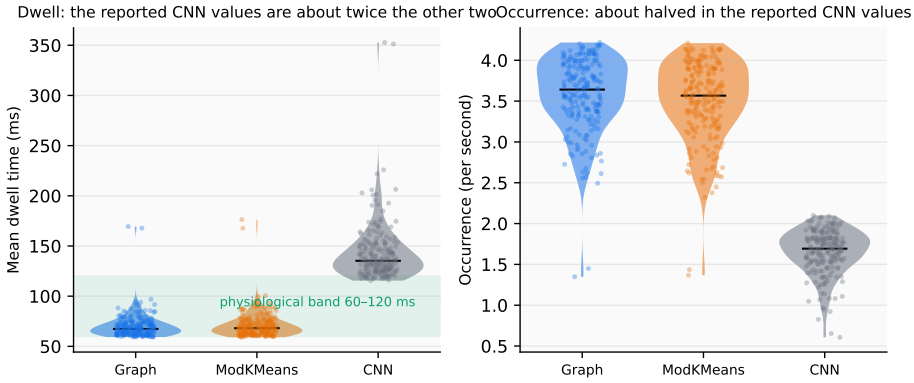

Figure 3: Per-participant temporal parameters at $K = 4$ with the classical reference. Top: dwell time; the shaded band marks the 60–120 ms physiological range. Bottom: occurrence rate. The graph-based model (blue) and modified K-means (orange) pace together inside the band; the CNN-based model (grey) sits outside it in the reported parameters.

Table 2: Paired comparison of the two models at K=4 (Wilcoxon signed-rank; p is Holm-adjusted over the 17 tested paired measures, with the equivalence rows that are identical by construction excluded from the family). The graph-based model carries the longer dwell, the higher occurrence, the higher backfit GEV, and the higher Calinski–Harabasz; the CNN-based model retains a significant silhouette edge and a lower Davies–Bouldin index.

Measure	Median		Mean		p_{Holm}	r_{rb}	n
	Graph	CNN	Graph	CNN			
GEV backfit (continuous)	0.616	0.491	0.618	0.506	$7.9e-34$	1.000	203
Silhouette	0.170	0.177	0.170	0.175	$1.5e-02$	-0.216	203
Mean dwell (ms)	67.357	135.239	70.243	143.207	$7.9e-34$	-1.000	203
Occurrence (/s)	3.641	1.693	3.568	1.655	$7.9e-34$	1.000	203
Entropy rate	0.398	0.215	0.393	0.210	$7.9e-34$	1.000	203
LZC (normalised)	0.204	0.107	0.201	0.105	$7.9e-34$	1.000	203
RQA latent z	56.695	24.397	58.069	24.707	$7.9e-34$	1.000	203

3.57 s^{-1} against 1.65 s^{-1}. Every one of the 203 pairs lies in the same direction at every K. Coverage of the smoothed backfit is complete for both models at $K = 4$ (100% of samples labelled), so the measure carries no separation at this operating point. Table 2 summarises the paired measures.

The temporal parameters depend on the decode. Figure 3 reports each model's parameters as its own pipeline produced them, and the two differ by a factor of two. Re-decoding each model's own recovered topographies with the same smoother at the same half-window and minimum segment removes the difference. Under that common decode the mean dwell is 70.2 ms for the graph-based model and 71.0 ms for the CNN-based model, and the occurrence rate is 3.57 against 3.37 s^{-1}. The residuals are 1% in dwell ($p = 7.7 \times 10^{-4}$) and 6% in occurrence ($p = 2.2 \times 10^{-27}$), detectable but an order of magnitude smaller than the reported divergence. Explained variance does not follow the same pattern. Under the common decode the graph-based model remains ahead, 0.618 against 0.542, in all 203 participants of the comparison ($p = 4.6 \times 10^{-35}$). The decode therefore accounts for the temporal divergence, and not for the difference in explained variance.

Internal cluster validity does not track the fit ranking. The CNN-based model attains the higher silhouette (0.175 against 0.170, $n = 203$, Holm-adjusted $p = 0.015$) and the lower Davies–Bouldin index. The graph-based model attains the higher Calinski–Harabasz index and the higher backfit GEV at every K. The indices are computed in different input geometries, so their scales are not comparable across the models, and the split is read by direction alone. No internal index tracks the validity ranking.

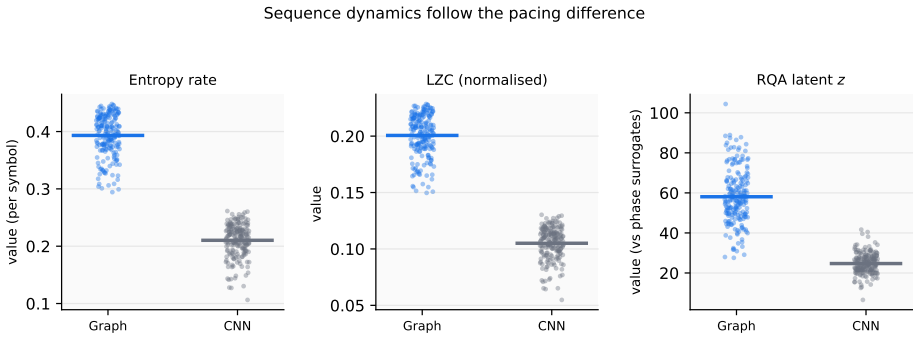

Figure 4: Sequence complexity and surrogate-normalised recurrence at $K = 4$. Each point is one participant; heavy lines are means. The graph-based model (blue) carries the higher entropy rate, the higher normalised Lempel–Ziv complexity, and the larger deviation from phase-randomised surrogates.

Figure 4 shows the sequence-level measures. Entropy rate is 0.393 against 0.210. Lempel–Ziv complexity is 0.201 against 0.105. The latent-trajectory recurrence deviates from phase-randomised surrogates by $z = 58.1$ against $z = 24.7$. All three differences survive Holm correction across the tested measures at $K = 4$ (Table 2). All three are computed on the label sequences, so a matched decode is required before their difference can be read as a property of the encoding; that control is reported above for dwell, occurrence, and explained variance. Transition statistics follow the same pattern as the maps rather than the dwell and occurrence. Under the common decode the CNN-based model shows the richer transition structure, with a median surrogate deviation of 12.7 against 8.3 for the graph-based model ($p = 2.5 \times 10^{-18}$) and a median absolute transition-matrix deviation of 13.9% against 8.3%. The graph-based model retains a marginal edge in the higher-order Markov statistic (2.94 against 2.91 on a $\log_{10}$ scale).

Table 3 extends the comparison to modified K-means under identical conditions. The classical reference dwells a median of 68.1 ms and falls inside the physiological band for 96.1% of participants; the graph-based model sits at 67.4 ms and 94.1%. On explained variance the reference attains a mean of 0.625 against 0.618 for the graph-based model, a margin of 0.7 points, while the CNN-based model sits 12 points below the reference. The reference is ahead on a paired test ($p = 1.3 \times 10^{-27}$), so the margin is small but systematic. On dwell the median of the paired differences is 0.45 ms ($p = 6.3 \times 10^{-10}$), and on occurrence it is -0.06 s^{-1} ($p = 4.3 \times 10^{-19}$); the two pace together while remaining distinguishable as distributions. The classical invariance assumption does not extend to the CNN-based model as its own pipeline reports it. The full set of 20 paired measures is given in Appendix A.

5 Discussion and Conclusion

The results separate what spatial encoding changes from what it does not. The reported temporal divergence between the two pipelines is a property of the decode. When both models' recovered topographies are decoded alike, dwell and occurrence agree to within a few per cent, and the classical reference paces with both. What survives the control is the difference in the recovered maps and in the transition statistics. Explained variance and agreement with the canonical templates (0.728 over 203 participants against a cohort summary of 0.400 over 196) both favour the graph encoding, and the explained-variance advantage is uniform across participants, while the transition statistics favour the CNN-based model. The comparison does not resolve to one model being better in every respect. Internal cluster metrics do not track it. Compactness in the input geometry is not validity of the recovered maps, and a significant silhouette edge can coexist with a lower explained variance. For the field, the consequence is a reporting obligation. Studies that report microstate parameters should state the model, the backfit, and the decode that produced them, and any comparison across pipelines should align the decode before a temporal claim is made. The confound is a property of pipeline configuration rather than of the encoding, and it is invisible to the metrics these models are usually evaluated with.

Table 3: Three-way comparison at $K=4$ ($n=203$). Dwell and occurrence are medians over participants, the in-band column is the share of participants whose mean dwell lies in the 60–120 ms physiological range, and GEV is the mean backfit global explained variance. Modified K-means paces with the graph-based model on all three measures; the CNN-based model is the outlier.

	Dwell (ms)	In band (%)	Occurrence (/s)	Backfit GEV
Graph-based model	67.4	94.1	3.64	0.618
Modified K-means	68.1	96.1	3.57	0.625
CNN-based model	135.2	7.9	1.69	0.506

The study has limitations. It examines one dataset, one state, and one montage, and the two pipelines are aligned on inputs by construction and by the in-data sameness check. Both models train on peaks of the same recording that is later backfit, so the reported fits are in-sample. Selection used one seed per run and a random peak-level split, and the architecture search that selected the graph configuration ran on ten participants drawn from the same cohort; blocked-split and multi-seed sensitivity are future work. The configurations also differ in latent dimension and loss terms, so the differences that survive the decode control are attributable to the pipeline configurations as a whole, with spatial encoding as the designed axis, rather than to the encoding operator alone. The rank-biserial effects saturate because no discordant pair exists in the explained-variance comparison, which the sameness check makes interpretable as a consistent front-end effect rather than an accident of one split. Test-retest reliability of microstate parameters remains an open concern (Kleinert et al., 2023) and is not assessed here. These limitations qualify, but do not change, the central finding: the reported temporal divergence between the two pipelines is a property of the decode, and the surviving differences between the encodings lie in the recovered maps and in the transition statistics, which do not favour the same model. Future work should test transfer across montages, test-retest reliability, and the same comparison in other modalities; the full set of paired measures, the configuration grid, and the metric definitions are given in the appendices.

AI-USE STATEMENT

AI tools were used in the preparation of this submission, with the main role being language editing, including grammar correction. All scientific content, claims, analyses, and decisions are the authors' own, matching the results of the analysis.

REPRODUCIBILITY STATEMENT

The dataset is publicly available at `https://ftp.gwdg.de/pub/misc/MPI-Leipzig_Mind-Brain-Body-LEMON/EEG_MPILMBB_LEMON/EEG_Preprocessed_BIDS_ID/EEG_Preprocessed/` and the cohort is its complete preprocessed archive. All models use deterministic seeds. The supplementary material contains the analysis code, the verification scripts, and the configuration files, and every number reported here is produced by one of those scripts rather than entered by hand.

REFERENCES

Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. Optuna: a next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 2623–2631, 2019. doi: 10.1145/3292500.3330701. arXiv:1907.10902.

Abdullah Almohammadi and Yu-Kai Wang. Revealing brain connectivity: graph embeddings for EEG representation learning and comparative analysis of structural and functional connectivity. *Frontiers in Neuroscience*, 17, 2024. doi: 10.3389/fnins.2023.1288433.

Aram Babayan, Miray Erbey, Deniz Kumral, Ronja Reinelt, Andrea M. F. Reiter, Anett Röbbig, Laura Lerebours, et al. A mind-brain-body dataset of MRI, EEG, cognition, emotion, and peripheral physiology in young and old adults. *Scientific Data*, 6:180308, 2019. doi: 10.1038/sdata.2018.308.

Adam P. Baker, Matthew J. Brookes, Isa A. Rezek, Stephen M. Smith, Timothy Behrens, P. J. Probert Smith, and Mark Woolrich. Fast transient networks in spontaneous human brain activity. *eLife*, 3:e01867, 2014. doi: 10.7554/eLife.01867.

Jelmer P. Borst and John R. Anderson. The discovery of processing stages: analyzing EEG data with hidden semi-Markov models. *NeuroImage*, 108:60–73, 2015. doi: 10.1016/j.neuroimage.2014.12.029.

Michaël Defferrard, Xavier Bresson, and Pierre Vandergheynst. Convolutional neural networks on graphs with fast localized spectral filtering. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2016. arXiv:1606.09375.

Saheed Faremi and Luca Longo. Integrating convolutional variational autoencoders and the Gaussian mixture model for efficient manifold learning and clustering of spatially preserved EEG topographic maps. *Brain Informatics*, 2026. doi: 10.1186/s40708-026-00327-9.

Saheed Faremi, Andrea Visentin, and Luca Longo. Interpretable EEG microstate discovery via variational deep embedding: a systematic architecture search with multi-quadrant evaluation. arXiv:2605.10947, 2026.

Alexandre Gramfort, Martin Luessi, Eric Larson, Denis A. Engemann, Christian Strohmeier, Christian Brodbeck, Roman Goj, Mainak Jas, Teon Brooks, Lauri Parkkonen, and Matti Hämäläinen. MEG and EEG data analysis with MNE-python. *Frontiers in Neuroscience*, 7:267, 2013. doi: 10.3389/fnins.2013.00267.

Nikola Jajcay and Jaroslav Hlinka. Towards a dynamical understanding of microstate analysis of M/EEG data. *NeuroImage*, 281:120371, 2023. doi: 10.1016/j.neuroimage.2023.120371.

Zhuxi Jiang, Yin Zheng, Huachun Tan, Bangsheng Tang, and Hui Zhou. Variational deep embedding: an unsupervised and generative approach to clustering. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 1965–1972, 2017. doi: 10.24963/ijcai.2017/273.

Arjun Khanna, Alvaro Pascual-Leone, Christoph M. Michel, and Faranak Farzan. Microstates in resting-state EEG: Current status and future directions. *Neuroscience & Biobehavioral Reviews*, 49:105–113, 2015. doi: 10.1016/j.neubiorev.2014.12.010.

Diederik P. Kingma and Max Welling. Auto-encoding variational Bayes. In *International Conference on Learning Representations (ICLR)*, 2014. arXiv:1312.6114.

Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations (ICLR)*, 2017. arXiv:1609.02907.

Timm Kleinert, Thomas Koenig, Kirsty Nash, and Edmund Wascher. On the reliability of the EEG microstate approach. *Brain Topography*, 37:271–286, 2023. doi: 10.1007/s10548-023-00982-9.

Dominik Klepl, Min Wu, and Fei He. Graph neural network-based EEG classification: a survey. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 32:493–503, 2024. doi: 10.1109/TNSRE.2024.3355750.

Thomas Koenig, Leslie Prichep, Dietrich Lehmann, Pedro Valdés Sosa, Elke Braeker, Hermann Kleinlogel, Ronald Isenhart, and E. Roy John. Millisecond by millisecond, year by year: normative EEG microstates and developmental stages. *NeuroImage*, 16(1):41–48, 2002. doi: 10.1006/nimg.2002.1070.

Dietrich Lehmann, Hiroshi Ozaki, and Ivo Pal. EEG alpha map series: brain micro-states by space-oriented adaptive segmentation. *Electroencephalography and Clinical Neurophysiology*, 67(3): 271–288, 1987. doi: 10.1016/0013-4694(87)90025-3.

Norbert Marwan, Marco Carmelo Romano, Marco Thiel, and Jürgen Kurths. Recurrence plots for the analysis of complex systems. *Physics Reports*, 438(5–6):237–329, 2007. doi: 10.1016/j.physrep.2006.11.001.

Christoph M. Michel and Thomas Koenig. EEG microstates as a tool for studying the temporal dynamics of whole-brain neuronal networks: a review. *NeuroImage*, 180:577–593, 2018. doi: 10.1016/j.neuroimage.2017.11.062.

Christoph M. Michel, Lucie Bréchet, Boris Schiller, and Thomas Koenig. Current state of EEG/ERP microstate research. *Brain Topography*, 37:169–180, 2024. doi: 10.1007/s10548-024-01037-3.

Michael Murphy, Hengyi Teresa Wang, Chunhui Jiang, Xiaoyi Wang, et al. A potential source of bias in group-level EEG microstate analysis. *Brain Topography*, 37:232–242, 2023. doi: 10.1007/s10548-023-00992-7.

Micah M. Murray, Denis Brunet, and Christoph M. Michel. Topographic ERP analyses: a step-by-step tutorial review. *Brain Topography*, 20(4):249–264, 2008. doi: 10.1007/s10548-008-0054-5.

Roberto D. Pascual-Marqui, Christoph M. Michel, and Dietrich Lehmann. Segmentation of brain electrical activity into microstates: model estimation and validation. *IEEE Transactions on Biomedical Engineering*, 42(7):658–665, 1995. doi: 10.1109/10.391164.

Aparna Sikka, H. Jamalabadi, Marina Krylova, et al. Investigating the temporal dynamics of electroencephalogram (EEG) microstates using recurrent neural networks. *Human Brain Mapping*, 41(9):2334–2346, 2020. doi: 10.1002/hbm.24949.

Sanat Thukral, Bujar Raufi, and Bojan Božic. Convolutional autoencoder-based dimensionality reduction for EEG microstate analysis. In *IFIP International Conference on Artificial Intelligence Applications and Innovations*, pp. 69–82, 2025. doi: 10.1007/978-3-031-96235-6_6.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph attention networks. In *International Conference on Learning Representations (ICLR)*, 2018. arXiv:1710.10903.

Frederic von Wegner, Alexander Knaut, and Helmut Laufs. EEG microstate sequences from different clustering algorithms are information-theoretically invariant. *Frontiers in Computational Neuroscience*, 12:70, 2018. doi: 10.3389/fncom.2018.00070.

Frederic von Wegner, Michael Wiemers, Gesine Hermann, Inken Tödt, Enzo Tagliazucchi, and Helmut Laufs. Complexity measures for EEG microstate sequences: concepts and algorithms. *Brain Topography*, 37:296–311, 2023. doi: 10.1007/s10548-023-01006-2.

Frederic von Wegner, Gesine Hermann, Inken Tödt, Inga Karin Todtenhaupt, and Helmut Laufs. Higher-order EEG microstate syntax and surrogate testing. *Computers in Biology and Medicine*, 193:110367, 2025. doi: 10.1016/j.compbiomed.2025.110367.

Junyuan Xie, Ross Girshick, and Ali Farhadi. Unsupervised deep embedding for clustering analysis. In *International Conference on Machine Learning (ICML)*, PMLR 48, pp. 478–487, 2016.

Table 4: Full set of paired measures at $K=4$ (21 measures). Per-measure n is 203 throughout. The equivalence symbol marks measures that are identical across the two models by construction (they depend only on the shared inputs), for which the paired test is undefined and which are excluded from the Holm family of 17 tested measures. The dagger marks a descriptive row, arithmetic on a design constant rather than a behavioural measure, which is printed for transparency and excluded from the family.

Measure	n	G med	C med	G mean	C mean	p_{Holm}	r_{rb}
GEV backfit (continuous)	203	0.6156	0.4914	0.6176	0.5063	7.9e-34	1
Silhouette	203	0.1702	0.1768	0.1697	0.1754	1.5e-02	-0.2164
Calinski-Harabasz	203	2540	257.8	2567	265.2	7.9e-34	1
Davies-Bouldin	203	1.589	1.457	1.613	1.504	2.7e-09	0.4957
Mean dwell (ms)	203	67.36	135.2	70.24	143.2	7.9e-34	-1
Occurrence (/s)	203	3.641	1.693	3.568	1.655	7.9e-34	1
Shannon entropy (full)	203	1.958	1.778	1.94	1.754	7.9e-34	0.9937
Entropy rate	203	0.3979	0.215	0.3932	0.2103	7.9e-34	1
LZC (normalised)	203	0.2036	0.1074	0.2006	0.105	7.9e-34	1
Mean Hurst	203	0.7929	0.8519	0.7946	0.8572	7.9e-34	-1
Markov p (order 1 vs 2)	203	1.4e-159	1.008e-22	7.161e-47	2.256e-05	7.9e-34	-1
RQA symbolic DET	203	0.9933	0.9985	0.9934	0.9985	7.9e-34	-1
RQA symbolic z (vs Markov)	203	2.785	1.423	2.668	1.337	7.9e-34	0.9999
RQA jump DET	203	0.5735	0.6451	0.5793	0.6533	1.1e-33	-0.9898
RQA jump z	203	0.02681	-0.1674	-0.08457	-0.1883	4.9e-02	0.1595
RQA GFP DET	203	0.7039	0.7039	0.6923	0.6923	$\equiv$	$\equiv$
RQA multivariate DET	203	0.8847	0.8847	0.8754	0.8754	$\equiv$	$\equiv$
RQA latent DET	203	0.8173	0.7088	0.8008	0.7036	1.1e-32	0.9729
RQA latent z	203	56.7	24.4	58.07	24.71	7.9e-34	1
Latent active dims	203	10	32	10	32	†	-1

A Full set of paired measures

B Configuration and training budgets

The reported cohort uses one fixed configuration, taken from the architecture search reported in the companion study (Faremi et al., 2026) and applied unchanged to every participant and every K. Training runs 100 pretraining epochs (learning rate 10^{-3}, batch 128, 20-epoch KL warm-up, patience 20) followed by 100 VaDE epochs (learning rate $5 \cdot 10^{-4}$, batch 128, patience 20), with K-means initialisation ($n_{\mathrm{init}} = 20$), a KL weight of 0.003, dropout 0.15 on the graph encoder, and deterministic seeds. A held-out fraction of 0.1 of the peaks is drawn at random and used for early stopping and for the validation metric. Every value is the one executed by the reported runs.

C Metric definitions

Entropy rate and Lempel–Ziv complexity follow the definitions and algorithms of von Wegner et al. (2023). Both use base-2 logarithms and are reported in bits, with the entropy rate expressed per symbol. Recurrence quantification follows Marwan et al. (2007) with parameters: window 1000 samples, embedding dimension 3, delay 10, target recurrence rate 0.05, minimal line length 2, Theiler window 1, and 99 surrogates (von Wegner et al., 2025), first-order Markov for the symbolic and jump sequences and phase-randomised, with a shared phase vector, for the continuous, multichannel, and latent representations. Long-range dependence is the detrended fluctuation analysis exponent of the per-state occupancy. The two GEV forms are defined in Section 3.

D Computational environment

Training ran on a Linux compute server under Python 3.10. The graph model is implemented in pure PyTorch (version 2.0.1) with NumPy; the graph convolution layers are implemented natively in the codebase and do not depend on external graph libraries. Preprocessing and evaluation use MNE-Python (Gramfort et al., 2013), SciPy, and scikit-learn.

The CNN comparator follows its published pipeline (Faremi & Longo, 2026), whose computational resources are documented in its own appendix.

The reported cohort comprises 3,654 graph-model runs (203 participants at each of 18 values of K), produced by the two-stage sweep and the automated search described in Appendix B. The classical anchor and the verification analyses run in Python with pandas and SciPy.

Training of the reported configuration ran on CPU workers (one thread per worker, deterministic seeds); no GPU is required for reproduction. The compute server is an IBM POWER9 system (ppc64le, little-endian) with two sockets of 20 cores at 2.3–3.8 GHz (160 hardware threads with four-way simultaneous multithreading) and six NUMA nodes, equipped with four NVIDIA Tesla V100-SXM2 16 GB GPUs.