Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Higher-order temporal structure in EEG microstate neurodynamics: an exact bigram-preserving surrogate test across three resting-state cohorts

Mehmet Berke Isler

mehmetberkeisler@hotmail.com

Istanbul Medipol University

Research Article

Keywords: EEG microstate sequences, Resting-state neurodynamics, Higher-order (non-Markovian) temporal dynamics, Surrogate data testing, Bigram-preserving surrogate, Symbolic dynamics

Posted Date: September 18th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10749583/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Higher-order temporal structure in EEG microstate neurodynamics: an exact bigram-preserving surrogate test across three resting-state cohorts

Mehmet Berke Isler[a,b,*]

[a]International School of Medicine, Istanbul Medipol University, Istanbul, Türkiye

[b]Department of Computer Engineering, School of Engineering and Natural Sciences, Istanbul Medipol University, Istanbul, Türkiye

[*]Corresponding author: mehmet.isler1@std.medipol.edu.tr

ORCID: 0009-0007-0822-6388

Abstract

Whether EEG microstate sequences carry temporal structure beyond a first-order (Markov-1) transition matrix bears on how memoryless large-scale brain dynamics are, but the answer depends critically on the surrogate that defines the null. Likelihood-based tests commonly compare a sequence model against a Markov-1 baseline using a shuffle that preserves only single-symbol (unigram) counts; this destroys the bigram statistics the baseline relies on and over-credits the model. We introduce an exact bigram-preserving surrogate (a randomised Eulerian-trail traversal of each subject's bigram multigraph) that reproduces the per-subject bigram count matrix *exactly* while resampling all higher-order structure, so a count-based Markov-1 likelihood is invariant under it by construction. On synthetic data the test held false positives at the nominal rate while remaining well powered. Across three independent resting-state cohorts (LEMON, Dortmund, SRM; $N = 931$), all processed identically, the corrected test isolated a small but reliably positive residual (its 95% confidence interval excluded zero in every cohort), while the naive unigram null over-credited the same detector by 1.1–1.8$\times$ (and a transformer up to threefold); two of the three cohorts agreed in magnitude ($+0.008$ and $+0.010$ bits/token), and the third was larger. The residual matches a generative first-order surrogate to within Monte-Carlo error and is captured as well by an order-2 Markov chain as by a transformer: small, short-range structure with a prominent second-order return tendency, not a neural-network discovery. We recommend exact bigram-preserving surrogates as a default conditional null when testing symbolic neural time series for structure beyond observed first-order transition counts.

Keywords EEG microstate sequences · Resting-state neurodynamics · Higher-order (non-Markovian) temporal dynamics · Surrogate data testing · Bigram-preserving surrogate · Symbolic dynamics

1 Introduction

Resting cortical activity is increasingly read as structured dynamics rather than noise: large-scale networks dwell in and switch between recurring, quasi-stable configurations, a metastable regime of coordination dynamics (Tognoli and Kelso 2014; Deco et al. 2011). The EEG microstate framework gives a

direct symbolic read-out of this process, segmenting resting EEG into a small set of quasi-stable scalp topographies (canonical maps A, B, C, D, ...) that index whole-brain network states and producing a discrete symbol stream, the natural unit of analysis for microstate dynamics (Michel and Koenig 2018; Pascual-Marqui et al. 1995). Classical analyses summarise this stream with first-order quantities: mean duration, occurrence, coverage, and the $K \times K$ transition matrix. How much of the temporal organisation of these metastable states a first-order (Markov) description captures, and how much reflects higher-order memory, is a recurring question in microstate neurodynamics: microstate sequences show scale-free, long-range temporal structure (Van de Ville et al. 2010; Gschwind et al. 2015), information-theoretic analyses reject low-order Markov models and quantify sequence entropy and complexity (von Wegner et al. 2017; von Wegner and Laufs 2018; Tarailis et al. 2025), a recurrent network bounded usable sequence memory at $\sim$1 s (Sikka et al. 2020), and a growing "microstate syntax" literature tests explicitly for structure beyond the transition matrix (Haydock et al. 2025).

A natural modern test is to train a sequence model and compare its held-out likelihood against a Markov-1 baseline. But a better likelihood does not by itself establish higher-order structure, for two reasons. First, a high-capacity model with a long context window can exploit each subject's overall map frequencies (information a Markov-1 table cannot represent) without using any sequence ordering. Second, and more subtly, the shuffled-surrogate control needs care. The most immediate permutation null, shuffling the token stream while preserving only single-symbol (unigram) counts, destroys the *bigram* statistics that the Markov-1 baseline itself depends on, so it handicaps Markov-1 and inflates the apparent advantage of any sequence model. Higher-order-syntax studies have instead compared against *generative* first-order Markov surrogates, which preserve the fitted transition probabilities in expectation (von Wegner et al. 2025, 2026); this is the right target, but it matches each subject's bigram statistics only on average, not exactly, leaving a residual mismatch in the very quantity the Markov-1 baseline reads off.

The correct control must preserve exactly what Markov-1 sees, the per-subject bigram (consecutive-pair) count matrix, while resampling all higher-order (trigram-and-longer) structure conditional on it. This is the *dinucleotide-preserving* shuffle, a standard tool in DNA and RNA sequence analysis (Altschul and Erickson 1985; Jiang et al. 2008); the same logic underlies degree-preserving rewiring nulls in network science (Maslov and Sneppen 2002), where the surrogate must preserve the low-order structure the test is meant to control for. To our knowledge it has not previously been applied to microstate sequences. The broader question, how much predictive structure a symbolic sequence carries beyond a finite-order Markov model, is the province of symbolic dynamics and computational mechanics (Shalizi and Crutchfield 2001); a prior pilot reconstructed microstate ε-machines in two single cases (Nehaniv and Antonova 2017), and our exact conditional surrogate offers a complementary, replication-oriented hypothesis test. We adapt the shuffle via a randomised Eulerian-trail (Hierholzer) construction and make three contributions: (i) building on prior demonstrations that generic shuffling can disrupt first-order microstate syntax and inflate higher-order inference (von Wegner et al. 2025, 2026), we make precise *why* a unigram-preserving shuffle is an invalid control for a Markov-1 comparison (it destroys the bigram statistics the baseline reads off) and quantify the inflation it produces; (ii) we give an exact bigram-preserving surrogate, show that the count-based Markov-1 likelihood is invariant under it, and validate on ground-truth synthetic data that it controls false positives at the nominal rate while remaining well powered; and (iii) across *three* independent resting-state cohorts we show that the corrected

control changes the conclusion, leaving a small but reliably positive residual that is captured as well by a classical higher-order Markov chain as by a transformer. Our contribution is therefore not the shuffle algorithm, which is classical, but the *exact conditional* test given each subject's observed bigram count matrix: we establish that the Markov-1 statistic is invariant under this surrogate, rather than matched only in expectation as under a generative null, which identifies it as the exact conditional null for structure beyond the observed first-order transition counts, and we quantify the surviving residual across three independent cohorts. The empirical effect is small, and we report it as such.

2 Materials and methods

Data. We used three independent public resting-state, eyes-closed EEG cohorts (Table 1). The primary cohort is LEMON (Babayan et al. 2019): 228 adults (20–77 yr), 62-channel BrainAmp, of whom 215 have a usable resting recording; after artefact-controlled preprocessing 213 yielded microstate token streams (two failed quality control), and all single-cohort analyses use these $n = 213$ subjects. For cross-cohort replication we added the Dortmund Vital Study (Getzmann et al. 2024) (OpenNeuro ds005385; 608 64-channel baseline recordings, all retained) and the SRM resting-state set (Hatlestad-Hall et al. 2022) (OpenNeuro ds003775; 64-channel, baseline session only, $n = 110$ unique individuals after one quality-control failure, with the test–retest sessions not pooled). LEMON's eyes-closed segments were extracted from the interleaved resting recording by its condition markers; the two OpenNeuro cohorts provide dedicated eyes-closed files.

Table 1: The three resting eyes-closed cohorts. Median tokens per subject and group GEV are from the identical K=5 pipeline; all analyses share the same 2–20 Hz preprocessing, tokenisation and surrogate.

Cohort	OpenNeuro	n	Channels	Median tokens/subj.	Group GEV
LEMON	–	213	62	7071	0.67
Dortmund	ds005385	608	64	1069	0.94
SRM	ds003775	110	64	3264	0.68

Preprocessing and microstate fitting. EEG was band-passed 2–20 Hz (the field-standard microstate band; Michel and Koenig 2018), notch-filtered, and resampled to 250 Hz. Non-scalp channels were dropped; bad channels were detected by local outlier factor (Breunig et al. 2000) and interpolated with spherical splines (Perrin et al. 1989); the data were average-referenced; and ocular, muscle, cardiac and line-noise independent components were removed automatically with ICLabel (Pion-Tonachini et al. 2019) in `MNE-Python` (Gramfort et al. 2013). To guard against the residual ocular/electrode artefacts that can masquerade as microstates, each extracted template was passed through a focality check, which all five $K = 5$ maps passed as broad cerebral topographies. Group-level modified K-means templates (Pascual-Marqui et al. 1995) were fit on global-field-power (GFP) peak maps with `pycrostates` (Férat et al. 2022) and back-fit per subject to yield a label sequence. We use $K = 5$ throughout, following the reliability analysis of Kleinert et al. (2023). Each stream is tokenised as a run-length-encoded sequence of either state-only symbols (the five microstate states) or (state, duration-bucket) symbols. Duration buckets use the log-spaced edges $[20, 40, 80, 160, 320, 640]$ ms (six edges, hence seven buckets), chosen to span the canonical microstate dwell range. This gives a *biological* alphabet of 5 (state-only) or $5 \times 7 = 35$ (state, duration) symbols; the models additionally reserve padding (PAD), sequence-start

(beginning-of-sequence, BOS) and sequence-end (EOS) tokens, so the *model* vocabularies are $V = 8$ and $V = 38$ respectively. We use the (state, duration) tokens throughout unless noted. Of the reserved tokens only the sequence-start marker actually enters these continuous streams, at position 0, contributing a single fixed start bigram that the surrogate preserves along with every other bigram; padding and end tokens never appear, so exact bigram preservation and Markov-1 invariance hold on the full model vocabulary.

Models. The test applies to any sequence model; we deliberately use two detectors against the same first-order baseline, a classical and a neural one, so the conclusion cannot rest on either. Our *primary* detector is a classical higher-order (Markov-2/3) chain: it is portable, GPU-free, directly targets second-order structure, and, as we show below, recovers the residual at least as strongly as the neural model; it is also the quantity we replicate across cohorts. As a secondary, high-capacity generalisation we use μMarkov-S, a small GPT-style causal transformer (Vaswani et al. 2017) ($d_{\text{model}} = 192$, 4 layers, 4 heads, weight-tied embeddings, pre-LayerNorm, 512-token context, $\sim$1.9 M parameters) trained by autoregressive next-token cross-entropy with AdamW (learning rate 3×10^{-4}, weight decay 0.01, cosine schedule, batch size 32, 20 epochs per fold) in PyTorch (Paszke et al. 2019), which shows the same surrogate principle transfers to a high-capacity model. The baseline is a Laplace-smoothed ($\alpha = 1$) first-order Markov (Markov-1) chain fit on the same tokens. Both are evaluated by held-out per-token negative log-likelihood (NLL, bits/token) under a 5-fold subject-wise split (Brookshire et al. 2024), so no test subject contributes to the *sequence model* it is scored by. (The group microstate templates are fit on all subjects, so the tokenisation itself is not cross-validated; this is standard for group microstate analysis and is orthogonal to the sequence structure we test, but our no-leakage guarantee is therefore for the sequence model, not the representation.) We report the paired per-subject gap $\text{NLL}_{M1} - \text{NLL}_{\mu}$ with a 95% percentile bootstrap CI over $B = 1000$ subject resamples. Given the small effect sizes, inference on the structure-specific comparisons is based on the confidence intervals rather than on p-values.

The bigram-preserving surrogate. The construction is illustrated in Fig. 1. For each subject we build the directed multigraph whose nodes are microstate tokens and whose edges are the observed consecutive pairs (bigrams), and take a randomised Eulerian trail through it (randomised Hierholzer construction; Hierholzer 1873). The trail traverses every edge exactly once, so the surrogate has an *identical* bigram count matrix while trigram-and-longer order is resampled. Such a trail always exists because the observed sequence is itself one: on its used edges the multigraph is connected, the start token has out-degree exactly one greater than its in-degree, the end token has in-degree exactly one greater than its out-degree, and every other node is balanced, precisely the Eulerian-*trail* condition (when the sequence starts and ends on the same token the trail is closed and all nodes are balanced). Algorithm 1 therefore always returns a valid surrogate of length N in $O(N)$ time. This is the classical dinucleotide shuffle (Altschul and Erickson 1985) applied to microstate tokens. Exact preservation of the bigram counts, which *every* trail shares, guarantees that the *Markov-1 baseline* is invariant under the surrogate. The distribution over trails is a separate matter and concerns only the *sequence model*: a well-specified test samples as uniformly as possible from the ensemble of sequences with the fixed bigram counts. Uniform sampling is provided by the Altschul-Erickson construction (which `uShuffle` realises through a random-arborescence step; Jiang et al. 2008), whereas the randomised Hierholzer traversal

we adopt for its $O(N)$ simplicity is not guaranteed to be uniform. We therefore verify empirically that the two samplers agree to within Monte-Carlo error on the structure-specific residual (Hierholzer $+0.00763$ versus uniform-Altschul-Erickson $+0.00753$ bits/token on LEMON, and agreeing to within Monte-Carlo error in all three cohorts; reported below), so our estimates are insensitive to sampler uniformity; the uniform construction is the principled default and is provided in the released code. We confirmed 100% bigram-count preservation on every subject.

The surrogate arm mirrors the real arm exactly: both μMarkov-S and Markov-1 are re-fit from scratch on the shuffled training folds and scored on the shuffled test folds under the identical 5-fold subject-wise split, so the structure-specific signal cannot reflect a train/test shift. Markov-1's NLL is unchanged by the surrogate to numerical precision, because it depends only on the preserved bigram counts. Writing C^{tr}_{ij} for the pooled bigram counts over the training subjects (from which the baseline is fit) and C^{te}_{ij} for those of a held-out subject s (on which it is scored), the cross-validated Markov-1 code length is

$$\mathrm{NLL}_{M1}(s) \;=\; -\sum_{i,j} C^{\mathrm{te}}_{ij} \log \hat{P}(j \mid i), \qquad \hat{P}(j \mid i) = \frac{C^{\mathrm{tr}}_{ij} + \alpha}{\sum_{j'} C^{\mathrm{tr}}_{ij'} + \alpha V}, \tag{1}$$

with $\alpha = 1$ (Laplace smoothing). Because the surrogate is applied per subject and preserves each subject's bigram counts exactly, it leaves every C^{tr}_{ij} (a sum over training subjects) and every C^{te}_{ij} unchanged; hence both the fitted $\hat{P}(j \mid i)$ and the scored code length are invariant, and the gap on shuffled tokens therefore measures only structure beyond first order. Both models score the same token positions (the initial BOS token is excluded), and the transformer's dropped final partial window cancels in the paired difference. The null hypothesis being tested is that, conditional on each subject's observed transition-count matrix, the stream carries no structure beyond first order, $\Pr(X_t \mid X_{<t}) = \Pr(X_t \mid X_{t-1})$; the bigram-preserving surrogate draws from exactly this conditional ensemble. Throughout, "exact" refers to the surrogate reproducing the observed bigram *count matrix* exactly, which is what makes the Markov-1 statistic invariant (Eq. 1); it does not assert *uniform* sampling within the ensemble. Uniform sampling is provided by the Altschul–Erickson construction, against which our efficient randomised-Hierholzer estimates are validated and found indistinguishable to Monte-Carlo error (reported below). For each subject we average the recomputed gap over three independent surrogates; the *structure-specific signal* is the per-subject paired difference (real $-$ averaged-surrogate gap), with a 95% percentile bootstrap CI over $B = 1000$ subject resamples, and we call the effect present when this CI excludes zero. The subject bootstrap resamples the fitted cross-validated per-subject scores; because subjects within a fold share a fitted model these scores are not fully independent, so the intervals are mildly optimistic (we return to this in the Limitations). For comparison we also report the looser unigram-preserving shuffle, which preserves only single-token counts.

Return-motif analysis. To characterise *what* the residual structure is, we computed on the state-only sequences (duration collapsed), for each subject, the frequency of each ordered trigram and its mean frequency over the bigram-preserving surrogates; the per-subject *return-motif index* is the excess frequency of *return* trigrams $X \to Y \to X$ (real minus surrogate), aggregated across subjects with the same paired bootstrap. This is a descriptive, exploratory analysis applied identically to all three cohorts.

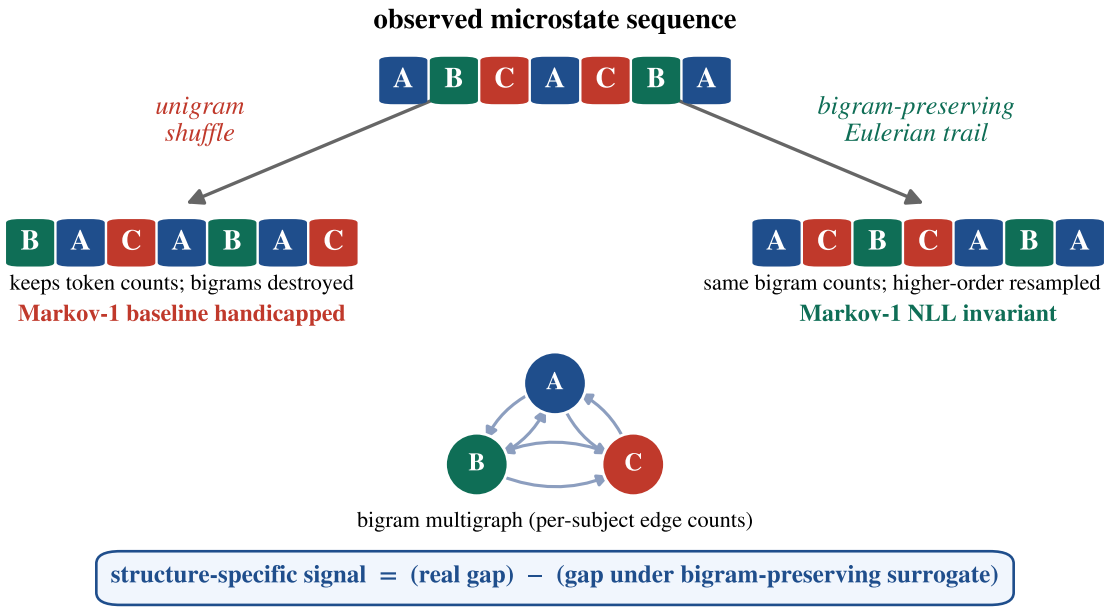


Fig. 1: The bigram-preserving surrogate, illustrated on a toy sequence (`A B C A C B A`; bigram multiset {AB, BC, CA, AC, CB, BA}). The conventional *unigram* shuffle (left) preserves single-token counts but destroys the bigram statistics the Markov-1 baseline relies on, handicapping the baseline. The *bigram-preserving* shuffle (right, `A C B C A B A`) is a randomised Eulerian trail through the subject's bigram multigraph: it reproduces the bigram multiset exactly while resampling trigram-and-longer order subject to that fixed multigraph, so the Markov-1 likelihood is invariant. The structure-specific signal is the real-minus-surrogate gap under this stricter control.

Algorithm 1 Exact bigram-preserving surrogate for a token sequence via a randomised Hierholzer Eulerian trail. Randomly ordering each vertex's out-edges yields a random Eulerian trail that reuses every observed bigram exactly once, so the bigram count matrix is preserved *exactly* (which is what leaves the Markov-1 likelihood invariant); this traversal does not, in general, sample uniformly over all compatible trails. The uniform Altschul-Erickson construction (Altschul and Erickson 1985), used as a sensitivity analysis and provided in the code, gives essentially identical estimates.

Require: token sequence $s = (s_1, \ldots, s_N)$
Ensure: surrogate s' with the same bigram count matrix as s

build directed multigraph G: for $i = 1, \ldots, N-1$, add edge $s_i \to s_{i+1}$
$start \leftarrow s_1$
for all vertices v of G **do**
    adj$[v] \leftarrow$ out-edges of v in **uniformly random** order
end for
$stack \leftarrow (start)$; $trail \leftarrow ()$
while *stack* is non-empty **do**
    $v \leftarrow \text{top}(stack)$
    if adj$[v]$ is non-empty **then**
        remove next edge $v \to w$ from adj$[v]$; push w onto *stack*
    else
        append pop$(stack)$ to *trail*
    end if
end while
return $s' \leftarrow \text{reverse}(trail)$ ▷ length N; each observed bigram used once

Implementation details. For reproducibility we record the key hyperparameters: a finite-impulse-response (FIR, `firwin`) band-pass and 50-Hz notch; bad channels flagged by local-outlier factor and spherical-spline interpolated *before* ICA; extended-Infomax ICA (99% variance retained, seed 42) with ICLabel (`mne-icalabel`; Li et al. 2022) removal at probability ≥ 0.80 for ocular, muscular, cardiac, line- and channel-noise components; modified K-means on GFP-peak maps (minimum peak distance 2 samples, 20 restarts, up to 10 000 peaks, polarity ignored), back-fit with a 5-sample smoothing half-window and a 32-ms minimum-segment constraint; and a focality gate (minimum cerebral fraction 0.15) that halts the pipeline if any map is focal rather than broad and cerebral. Configuration files, random seeds, and per-fold subject-ID splits ship with the code. The analysis ran under Python 3.10 with MNE-Python 1.8.0, `pycrostates` 0.6.1, `mne-icalabel` 0.9.0, scikit-learn 1.5.0, NumPy 1.26.4, SciPy 1.13.1 and PyTorch 2.4.1.

3 Results

Validation on synthetic data. Before applying the test to EEG, we verified its operating characteristics on synthetic symbolic sequences with known ground truth (Fig. 2). The detector was a classical Laplace Markov model (a first-order baseline against an order-k model), scored by the same 5-fold subject-wise paired procedure used throughout; because the bigram-preserving surrogate leaves the Markov-1 likelihood invariant, this isolates the surrogate's behaviour independently of the detector. Sequences were run-length-encoded microstate-like streams ($V = 5$, no self-transitions), 150 simulated subjects of realistic length ($\sim$1200 tokens), with 200 replications per null condition and 80 per alternative. *False positives.* On sequences from a true first-order chain (no structure beyond first order), the bigram-preserving surrogate held the false-positive rate at 2.5%, consistent with the nominal 5% given the 200 replications (binomial SE $\approx 1.5\%$), whereas the conventional unigram-preserving shuffle inflated it to 12% (Fig. 2, left), the failure mode described in the Introduction, now demonstrated on ground truth. *Power.* On sequences with a planted second-order dependency of increasing strength λ, the bigram-preserving structure-specific signal rose monotonically from zero and tracked λ, with detection power reaching 1.0 for even weak departures (Fig. 2, right); a planted third-order dependency was likewise detected (signal $+0.65$ bits/token with an order-3 detector). *Subject heterogeneity.* When each simulated subject was given its own first-order chain (heterogeneous unigram and bigram distributions but no higher-order structure), the false-positive rate remained at 1.5%, confirming that per-subject heterogeneity is not mistaken for higher-order syntax. The test is therefore correctly calibrated and adequately powered, and the unigram control is empirically miscalibrated under a structured null. Two further checks confirm the procedure is insensitive to its own tuning. The number of surrogates averaged per subject changes only the Monte-Carlo noise on the surrogate mean, not the expected residual, so the three averaged throughout suffice. So is the trail sampler, which matters for the sequence model even though the Markov-1 baseline is invariant either way: replacing the randomised Hierholzer traversal with a uniform Altschul-Erickson construction (a random-arborescence sampler, as in `uShuffle`; Jiang et al. 2008) gives an essentially identical residual in every cohort (LEMON $+0.00753$ vs. Hierholzer $+0.00763$; Dortmund $+0.00944$ vs. $+0.00952$; SRM $+0.03176$ vs. $+0.03183$), showing empirically that the non-uniformity of the $O(N)$ traversal does not bias the estimate. Finally, repeating the null and Markov-2 conditions with the actual transformer detector reproduced the key property: a ≈ 0 signal

with no false rejections on first-order data and a positive signal on second-order data, so the surrogate controls false positives and retains power for the model the paper uses.

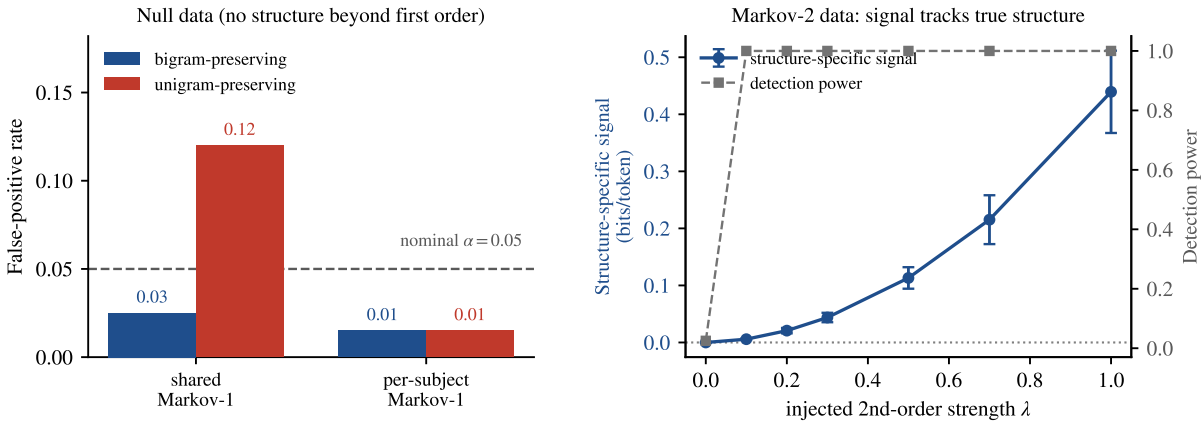


Fig. 2: Ground-truth validation of the surrogate test (classical Markov detector, 5-fold subject-wise paired procedure; 150 simulated subjects). *Left:* on data with no structure beyond first order, the bigram-preserving surrogate holds the false-positive rate at the nominal level (2.5% under a shared first-order chain, 1.5% under per-subject heterogeneous chains), whereas the unigram-preserving shuffle inflates it to 12% under the shared-chain null. *Right:* on Markov-2 data with injected second-order strength λ, the bigram-preserving structure-specific signal rises monotonically from zero (mean $\pm$ SD across replications) and detection power reaches 1.0, so the test is well powered for genuine higher-order structure.

Microstate fit. The four canonical maps (A, B, C, D) plus an auxiliary fifth map were recovered at $K = 5$; all five are broad, cerebral topographies with no focal ocular/electrode signature, and are shown in Fig. 3. Fit quality is reported as the global explained variance at GFP peaks (GEV, the fraction of GFP-peak map variance explained by the assigned templates, computed at the GFP peaks before any temporal smoothing): 0.67 for LEMON, 0.94 for Dortmund and 0.68 for SRM. All three use the identical peaks-based estimator and the identical preprocessing pipeline; the same formula is applied throughout, so the values are directly comparable. Dortmund's higher GEV coincides with its much shorter per-subject recordings (Table 1: a median of $\sim$1,000 tokens versus $\sim$7,000 for LEMON); the reason for this unusually high value is not established, but we verified it is not an artefact of a different GEV definition or preprocessing (all three cohorts use the identical peaks-based estimator, computed before temporal smoothing, under the identical pipeline). Crucially, whatever its origin, fit quality does not set the residual: LEMON and SRM have near-identical GEV (0.67 vs. 0.68) yet a fourfold difference in residual, and *within* each cohort the per-subject GEV correlates *positively* with the residual (Section 3.1), the opposite of what a low-fit artefact would produce.

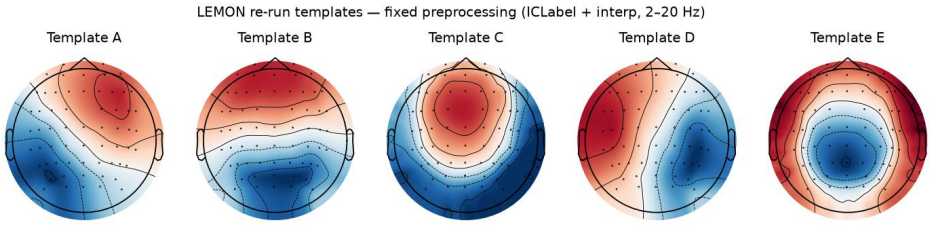


Fig. 3: LEMON modified-K-means microstate templates at K=5 ($n = 213$; artefact-cleaned preprocessing: ICLabel + bad-channel interpolation, 2–20 Hz). All five maps are broad cerebral topographies. Back-fitting used competitive smoothing (5-sample half-window) with a 32-ms minimum-segment constraint; maps are polarity-invariant in analysis and polarity-canonicalised only for visualisation.

Raw gap. At $K = 5$, on the (state, duration-bucket) vocabulary ($V = 38$), the adjudicating quantity is the *subject-paired* gap $\mathrm{NLL}_{M1} - \mathrm{NLL}_{\mu}$, which weights each subject equally: it is $+0.035$ bits/token (95% CI $[+0.025, +0.051]$, $n = 213$). Taken at face value this looks like higher-order structure, but most of it is not.

The two surrogates. The structure-specific signals under both surrogates are summarised in Table 2 and shown as paired bootstrap intervals in Fig. 4. Under the *unigram*-preserving shuffle the gap on shuffled tokens drops to a floor of $+0.017$ bits/token, so this control attributes a large structure-specific signal of $+0.018$ bits/token (95% CI $[+0.014, +0.023]$, $n = 213$) to the model. But this is an overestimate: the unigram shuffle also destroys the bigram statistics Markov-1 relies on, handicapping the baseline and over-crediting the model. This is the failure mode the paper is about.

Table 2: Key quantities at K=5. The raw gap and the gaps under each surrogate are the Markov-1 $-$ μMarkov-S NLL gaps on real and shuffled tokens; structure-specific signals are paired per-subject differences (real $-$ surrogate) with 95% bootstrap CIs. The bigram-preserving surrogate is the adjudicating control.

Quantity	Value	95% CI	n
Raw gap (real)	$+0.035$	$[+0.025, +0.051]$	213
Gap under unigram surrogate	$+0.017$	–	213
Gap under bigram surrogate	$+0.029$	–	213
Struct.-specific, unigram surrogate	$+0.018$	$[+0.014, +0.023]$	213
Struct.-specific, bigram surrogate	$\mathbf{+0.0059}$	$[\mathbf{+0.0050}, \mathbf{+0.0068}]$	**213**

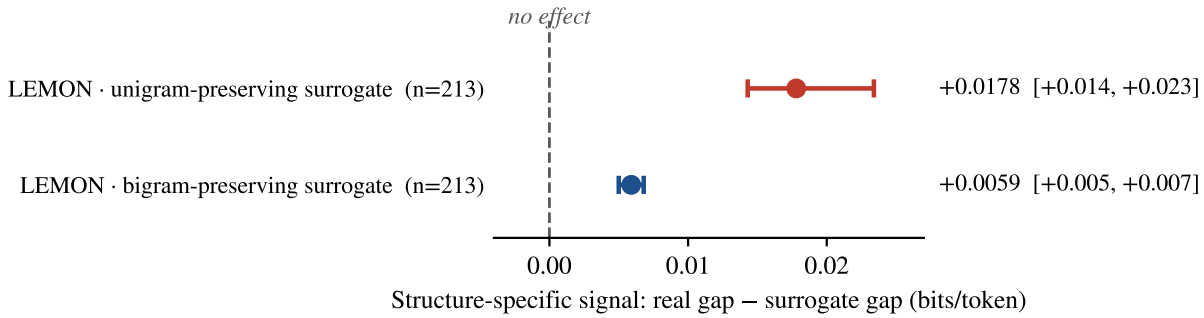


Fig. 4: Structure-specific signals (paired per-subject difference, real gap minus surrogate gap; $n = 213$) with 95% bootstrap confidence intervals. The conventional unigram-preserving surrogate over-credits the model ($+0.018$ $[+0.014, +0.023]$); the stricter bigram-preserving surrogate, which keeps the bigram statistics Markov-1 relies on, isolates a much smaller correctly controlled residual ($+0.0059$ $[+0.0050, +0.0068]$). Both exclude zero, but the unigram estimate is inflated roughly threefold.

Under the stricter bigram-preserving surrogate, Markov-1 keeps its bigram statistics, and the gap on shuffled tokens drops only to $+0.029$ bits/token, so per-subject bigram context accounts for most of the raw $+0.035$ gap, leaving a residual of $+0.0059$ bits/token (Fig. 5). The matched paired structure-specific signal is $+0.0059$ bits/token (95% CI $[+0.0050, +0.0068]$, $n = 213$): a small but reliably positive residual whose CI *excludes zero*. To fix intuition, $+0.0059$ bits/token is about 0.3% of the $\sim$2-bit per-symbol next-state uncertainty at K=5, small in absolute terms but consistently signed across subjects and cohorts. The unigram control puts this residual at $+0.018$, roughly threefold larger, because it handicaps Markov-1; the bigram-preserving surrogate gives the correctly controlled value (Fig. 4). Because this residual is small in absolute terms, we characterise its stability directly below: it survives re-fitting the

microstate templates from independent seeds, changing the number of maps K, and removing all duration information, and it is baseline-independent by construction rather than an artefact of how the first-order baseline is smoothed.

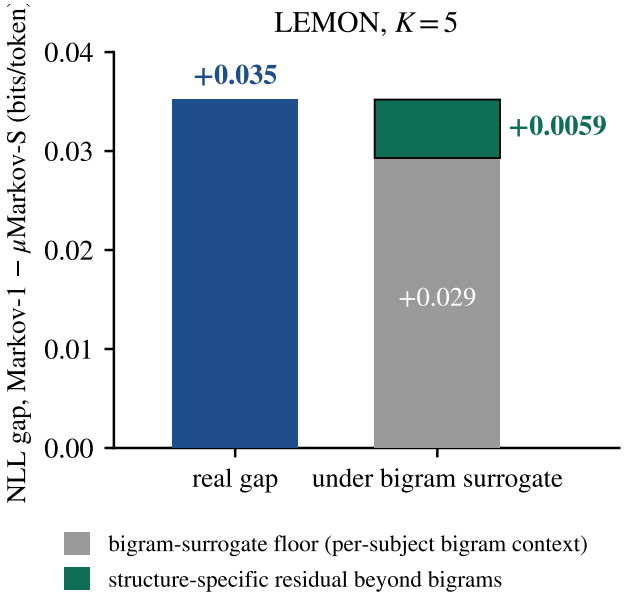


Fig. 5: Decomposition of the raw LEMON K=5 gap ($n = 213$). The bigram-preserving surrogate reproduces $+0.029$ of the $+0.035$ raw gap as per-subject bigram context (grey); only the $+0.0059$ residual (teal) survives as structure beyond the bigram counts.

Robustness. The residual is small, so we checked it against the analysis choices most likely to inflate it (Table 3, all $n = 213$). *Template-fitting seed.* Re-fitting the ModKMeans maps from three independent random seeds gives structure-specific residuals of $+0.0058$, $+0.0043$ and $+0.0044$ bits/token (mean $+0.0048$, SD 0.0007 across seeds); every seed's CI excludes zero, so the effect is not an artefact of one template solution. *Number of maps.* Sweeping $K \in \{4,6,7\}$ gives $+0.0044$, $+0.0052$ and $+0.0057$ bits/token, all CIs excluding zero, so the residual does not depend on K=5. *Duration versus syntax.* On the state-only sequence with duration removed (V=5) the residual is $+0.0030$ bits/token ($[+0.0023, +0.0040]$), so about half of the V=38 residual is genuine microstate *syntax* beyond the bigram counts and roughly half is duration dynamics; the two components are comparable, and we do not claim syntax dominates. *The baseline cancels exactly.* Because the count-based (Laplace) Markov-1 held-out likelihood is a function only of the preserved bigram counts, the fixed vocabulary and the preserved start token, it is invariant under the surrogate to machine precision (mean $|\Delta| \approx 5 \times 10^{-15}$ bits/token across subjects). The structure-specific signal therefore reduces *identically* to the sequence model's own real-versus-surrogate difference, $\mathrm{NLL}_\mu(\text{surrogate}) - \mathrm{NLL}_\mu(\text{real})$: the Markov-1 term does no arithmetic work in the adjudicating quantity. This makes the signal immune to how the baseline is smoothed (Laplace versus absolute-discounting), but it also means the invariance is a property of the count/Laplace Markov-1 specifically, not of an arbitrary first-order model.

Table 3: Robustness of the structure-specific residual (paired, real − bigram-surrogate gap; bits/token; 95% bootstrap CI; $n = 213$). The template-seed, number-of-maps and duration-removed rows use the transformer detector; the two *Classical* rows use classical Markov chains on the identical V=38 tokens. The residual is positive with a CI excluding zero under every perturbation, and the classical chains capture it at least as strongly as the transformer. These cells re-analyse the same $n = 213$ subjects, so they are correlated robustness checks, not independent replications (for which see the cross-cohort analysis).

Perturbation	Variant	Residual	95% CI
Template seed	0	+0.0058	[+0.0049, +0.0068]
	1	+0.0043	[+0.0033, +0.0053]
	2	+0.0044	[+0.0032, +0.0054]
	mean (SD 0.0007)	+0.0048	–
Number of maps K	4	+0.0044	[+0.0035, +0.0056]
	6	+0.0052	[+0.0043, +0.0064]
	7	+0.0057	[+0.0047, +0.0066]
Duration removed	state-only (V=5)	+0.0030	[+0.0023, +0.0040]
Classical, same V=38	Markov-2	+0.0076	[+0.0062, +0.0094]
	Markov-3	+0.0105	[+0.0090, +0.0121]

A count model, not the transformer, is the primary detector. Because it is portable, GPU-free, and, as we now show, at least as sensitive, we take the classical Markov-2 structure-specific residual as the primary estimand from here on (it is also the quantity replicated across cohorts below); the transformer is reported as a concordant secondary check. Computing the *same* structure-specific signal for classical higher-order Markov chains on the *identical* V=38 tokens, under the same bigram-preserving surrogate (Table 3, classical rows), a second-order chain (Laplace-smoothed, matching the Markov-1 baseline) yields +0.0076 bits/token ([+0.0062, +0.0094]) and a third-order chain +0.0105 ([+0.0090, +0.0121]), both slightly *larger* than the small transformer's +0.0059 on the same tokens, consistent with a 1.9 M-parameter model being under-powered on a corpus this size. The beyond-bigram structure is thus captured at least as well by a two-count-table model as by a neural network. Its predictability is also short-range: even the third-order chain (∼150–250 ms of context) and the 512-token transformer add only about 0.006 bits/token over Markov-1 (held-out NLL 1.998 → 1.992 from Markov-1 to an interpolated absolute-discounting Markov-4 in the sense of Kneser and Ney 1995). We therefore describe the extractable structure as small and short-range; this does not by itself adjudicate scale-free or long-range *autocorrelation*, a different construct we do not test.

No behavioural anchor. An exploratory check found that the per-subject residual does not track chronological age (Spearman $\rho = -0.12$, $p = 0.08$, $n = 213$) and adds no cross-validated predictive value over the four classical microstate features, so we make no biomarker claim.

3.1 Replication across independent cohorts

The LEMON residual is small, so a single cohort cannot establish that it is a general property of resting microstate sequences rather than an idiosyncrasy of one acquisition. We therefore applied the *same* portable detector, the classical Markov-2 structure-specific test (Table 3, bottom), which needs no GPU and no per-cohort tuning, to the two further cohorts described in the Data section (Dortmund and SRM; Table 1), fitting K=5 microstate templates *independently* in each (templates are never transferred). Preprocessing (*including the identical 2–20 Hz band, 30 ICLabel components and random seed*), tokeni-

sation (V=38), the bigram-preserving surrogate, and the paired bootstrap were identical to LEMON; each cohort's K=5 templates were broad cerebral maps that passed the focality gate and are shown in Supplementary Fig. S1.

Two conclusions replicate in every cohort (Table 4, Fig. 6). First, the bigram-preserving residual is small but *reliably positive*: its 95% bootstrap CI, percentile and bias-corrected-and-accelerated (BCa) alike, excludes zero in all three cohorts, and it is positive in the large majority of individual subjects (LEMON 87%, Dortmund 84%, SRM 87%; Fig. 7). Second, the naive unigram-preserving control *over-credits* the residual in every cohort, by 1.1–1.8× with this common classical detector (LEMON 1.3×, Dortmund 1.8×, SRM 1.1×). A more flexible model is over-credited more: on LEMON the same naive control inflates the small transformer's residual roughly threefold (Table 2), so the naive surrogate is most dangerous exactly where a spurious "discovery" is most tempting.

With preprocessing matched across cohorts, two of the three now *agree in magnitude*: LEMON (+0.0076) and Dortmund (+0.0095) are nearly identical, while SRM (+0.0318) is three- to four-fold larger. This difference is not an artefact of the analysis choices most likely to produce it. It is not a sequence-length effect: truncating every subject to a common token budget (SRM records $\sim 3\times$ more tokens per subject than Dortmund) leaves SRM's residual as large as ever (+0.035 to +0.037 across truncation lengths, if anything slightly above the untruncated +0.0318, all 110 SRM subjects retained at every length), so its magnitude is not an artefact of long recordings and is not reduced by matching Dortmund's sequence length. Nor is it manufactured by the back-fitting smoothing: on Dortmund the residual *survives* removing the temporal smoothing and minimum-segment constraint entirely (rising to +0.019), so the smoothing, if anything, makes the test more conservative, though the absolute magnitude does depend on these choices, which we flag as a caveat. It is not a fit-quality artefact either: there is no evidence that the larger residual is driven by poorer microstate fit. The per-subject residual does not increase as fit worsens; if anything it correlates *positively* with that subject's microstate GEV within every cohort (Spearman $\rho = +0.12$ to $+0.79$, all $p < 0.01$). And between cohorts fit quality is decoupled from magnitude: LEMON and SRM have near-identical GEV (0.67 and 0.68) yet residuals differing four-fold. Finally, it does not depend on the duration binning: re-tokenising with shifted, finer, or data-driven quantile bucket edges leaves each cohort's residual within its bootstrap interval (LEMON +0.007 to +0.008, Dortmund +0.009 to +0.010, SRM +0.030 to +0.032 across all four bin sets). We therefore read the cross-cohort picture as a magnitude replication between LEMON and Dortmund and a larger residual in SRM that is robust within our pipeline (to sequence length, smoothing, duration binning and fit quality) but whose biological origin we have not established, rather than one universal effect size; the *direction* of the effect and the over-crediting of the naive surrogate replicate everywhere.

Table 4: Cross-cohort replication of the classical Markov-2 structure-specific residual (bits/token; 95% bootstrap CI) under the bigram-preserving surrogate. Templates were fit independently per cohort at K=5. The residual is positive with a CI excluding zero in every cohort, and the unigram-preserving control over-credits it in every cohort; the *magnitude*, however, is cohort-specific (see text), so we report each cohort separately rather than a pooled estimate.

Cohort	OpenNeuro	n	Bigram residual [95% CI]	Unigram
LEMON	–	213	$+0.0076$ $[+0.0062, +0.0094]$	$+0.0096$ (1.3×)
Dortmund	ds005385	608	$+0.0095$ $[+0.0086, +0.0105]$	$+0.0169$ (1.8×)
SRM	ds003775	110	$+0.0318$ $[+0.0265, +0.0371]$	$+0.0362$ (1.1×)

All three cohorts use the identical 2–20 Hz pipeline and the same classical Markov-2 detector, so the bigram and unigram columns are directly comparable across cohorts. LEMON and Dortmund agree in magnitude, while SRM is larger (robust to sequence length and smoothing; see text); the naive unigram control over-credits the residual by 1.1–1.8× in every cohort.

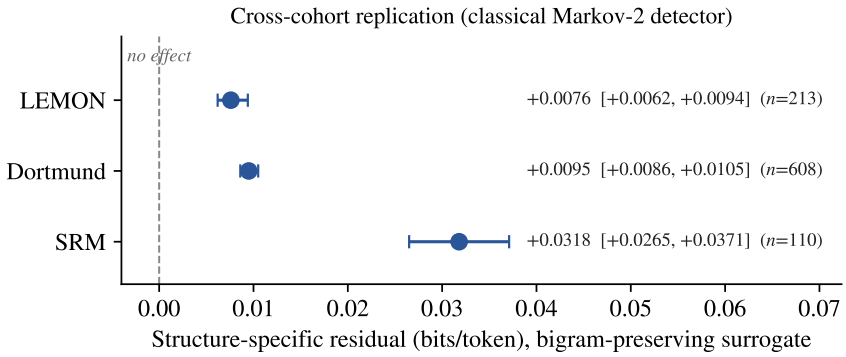


Fig. 6: Cross-cohort forest plot of the classical Markov-2 structure-specific residual under the bigram-preserving surrogate (bits/token; 95% bootstrap CI). The residual excludes zero in every cohort, so the *direction* of the effect replicates; the *magnitude* is heterogeneous and cohort-specific, so no single pooled estimate is reported.

Inference robustness. The headline intervals are 95% percentile bootstrap CIs that resample *subjects* with replacement (equal weight per subject, not token-weighted), the natural unit of independence. Because these resample the fitted cross-validated scores while holding the fitted models fixed, we also ran a leakage-free grouped (cluster) bootstrap that refits the models inside each resample: it assigns every copy of a resampled subject to a single cross-validation fold and scores only held-out subjects, so no subject appears in both training and test. Over $B = 300$ resamples per cohort the refit intervals essentially reproduce the headline ones and exclude zero in all three (LEMON $+0.0076$, $[+0.0062, +0.0094]$; Dortmund $+0.0093$, $[+0.0081, +0.0106]$; SRM $+0.0315$, $[+0.0254, +0.0385]$), confirming the effect is not an artefact of treating the fitted models as fixed. We do not report a pooled cross-cohort effect: with three cohorts and near-complete between-cohort heterogeneity ($I^2 \approx 97\%$) no common effect size is defensible, so the shared exclusion of zero, not a meta-analytic point estimate, carries the replication claim.

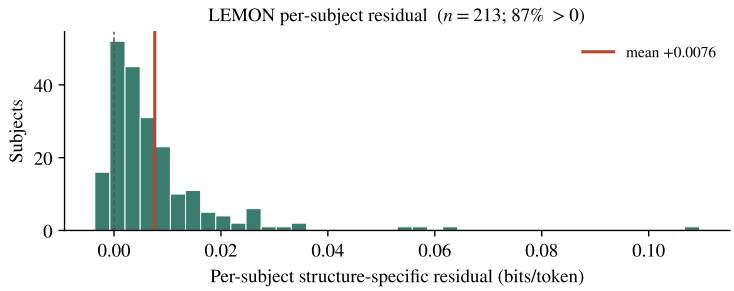


Fig. 7: Per-subject structure-specific residual on LEMON ($n = 213$; classical Markov-2 detector under the bigram-preserving surrogate, three surrogates averaged). The distribution is positive for 87% of subjects (median +0.0045, mean +0.0076 bits/token); the small residual is a broad across-subject tendency, not an artefact of a few outliers.

3.2 Trigram structure: a reproducible return tendency

Having established that a small higher-order residual survives, we asked what it is. On the state-only sequences (duration collapsed) we computed, for each subject, a *return-motif index*: the excess frequency of *return* trigrams $X \to Y \to X$ (the state preceding an excursion recurring after it) over their mean frequency under the exact bigram-preserving null. This is a descriptive characterisation, computed the same way in all three cohorts rather than a pre-registered hypothesis discovered in one and tested in the others. The return excess is positive with a bootstrap CI excluding zero in every cohort, and positive in the large majority of individual subjects (LEMON $+0.023$ $[+0.021, +0.025]$, 99% of subjects; Dortmund $+0.011$ $[+0.010, +0.012]$, 84%; SRM $+0.058$ $[+0.052, +0.066]$, 97%; Fig. 8), while the complementary *transit* trigrams $X \to Y \to Z$ ($Z \neq X$) are correspondingly under-represented. The beyond-bigram structure is therefore a mild non-Markovian *return* tendency that a first-order transition matrix and its Eulerian-trail null cannot represent. (Which intermediate map is traversed differs between cohorts, as expected when templates are labelled independently; the return *geometry* is the consistent feature.)

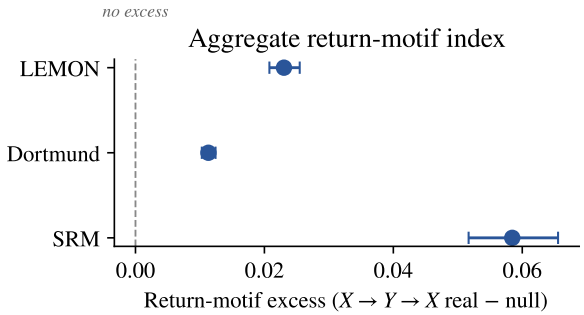


Fig. 8: Aggregate return-motif index per cohort (per-subject excess frequency of $X \to Y \to X$ return trigrams over the exact bigram-preserving null; mean and 95% bootstrap CI). The excess excludes zero in all three cohorts.

4 Discussion

The methodological point is simple and, we think, general. When a sequence model is compared against a Markov-1 baseline on microstate tokens, the control used to certify "this is higher-order structure" must preserve the bigram statistics the baseline relies on. The conventional unigram-preserving shuffle does not: it destroys those bigrams, handicaps Markov-1, and so over-credits the sequence

model. On LEMON this matters quantitatively: the unigram control reports a structure-specific signal of $+0.018$ bits/token, roughly threefold larger than the $+0.0059$ the bigram-preserving surrogate isolates once Markov-1 keeps its bigram statistics. The two surrogates give quantitatively different conclusions from the same data, and the bigram-preserving one is the exact conditional null for structure beyond the observed first-order transition counts.

This connects directly to recent work on higher-order microstate syntax, which tests for structure beyond the first-order transition matrix using *generative* first-order Markov surrogates or entropy-rate statistics (von Wegner et al. 2025, 2026). Those surrogates preserve the fitted transition *probabilities* in expectation; our Eulerian-trail surrogate preserves each subject's observed bigram *count matrix* exactly on every surrogate, so the Markov-1 likelihood is invariant by construction rather than in expectation, and the structure-specific signal needs no null-model fitting variance. Crucially, this is a refinement, not a reversal: on the LEMON state sequences the generative first-order surrogate and the exact bigram-preserving surrogate give structure-specific residuals that match to within Monte-Carlo error, whereas the naive unigram null inflates the same quantity by roughly one-half. The exact surrogate's advantage over the generative one is therefore the elimination of null-model fitting variance, not a different answer; the inflation we warn against is specific to the naive unigram/permutation null used in some likelihood-based tests, not to the generative-surrogate line of work, which our exact-count Eulerian surrogate complements. The bigram-preserving surrogate applies to any tokenised state stream, including HMM-derived state sequences (Vidaurre et al. 2017).

On the substantive side we are deliberately modest. The primary (classical Markov-2) residual on LEMON is $+0.0076$ bits/token (about 0.4% of the $\sim$2-bit next-symbol uncertainty), and the small transformer recovers slightly less, so the effect is not a neural-network "discovery": the beyond-bigram structure is captured at least as well by a two-count-table model as by a network, and no neural machinery is required to detect or interpret it. The residual is not carried by a handful of subjects: it is positive for 87% of them (Fig. 7), though the per-subject distribution is right-skewed (median $+0.004$ below the mean $+0.008$ bits/token), so a minority contribute disproportionately to the mean even as the positive sign is shared broadly. It carries no behavioural anchor we could detect: the age association is fragile and adds nothing over classical features. The honest reading is that microstate sequences contain a little genuine structure beyond a bigram model, and that prior sequence-model advantages over Markov-1 are largely explained by per-subject context once the exact conditional null is used.

In neurodynamic terms, the residual quantifies the memory carried by the symbolic trajectory of large-scale EEG topographies once the first-order transition geometry has been fixed. Microstates coarse-grain metastable cortical dynamics into a small symbolic alphabet (Kelso 2012; Tognoli and Kelso 2014), so testing for structure beyond the observed transition counts asks how much of the sequence's predictability reflects short temporal context rather than the stationary jump probabilities between states. The answer here is: a little, and it is short-range. The reproducible $X \to Y \to X$ return excess points to a mild recurrence tendency (after a brief excursion the system is slightly more likely to revisit the immediately preceding topography than its transition-count matrix alone predicts), which is consistent with microstates as a symbolic projection of metastable dynamics whose dwell-and-return geometry leaves a faint second-order trace. One possible dynamical interpretation, which our scalp-level analysis does not establish, is that such a return tendency reflects the transient, heteroclinic-channel organisation posited for cortical dynamics, in which the system is drawn back toward recently visited

states (Rabinovich et al. 2008); it is also consistent with hidden-Markov evidence for fast, recurring transient network states in spontaneous activity (Baker et al. 2014). The very small magnitude, however, argues against a rich long-context grammar in resting-state microstate streams: the beyond-bigram memory is real but shallow, dominated by second-order recurrence rather than long-range syntax, in line with the modest sequence complexity and predictability reported for resting microstates (Tarailis et al. 2025; Jahromy 2025). Read this way, the residual is best understood as an empirical upper bound on the higher-order memory that a symbolic microstate model can add over first-order transition counts in resting-state cortex, not as evidence for long-range non-Markovian dynamics.

Limitations. The three cohorts are all healthy adults recorded at rest with eyes closed; transfer to clinical or paediatric data is untested. The residual's magnitude replicated between LEMON and Dortmund but was three- to four-fold larger in SRM; although we ruled out sequence length and back-fitting smoothing as explanations, we did not identify which population or acquisition factor drives SRM's larger value, and the magnitude does depend on the back-fitting smoothing, so absolute magnitudes should be read with caution. Ages are bucketed, and the age analysis does not control for sex. The headline interval is a single-template bootstrap, but the residual is stable across independent template seeds and across K (Table 3), so this does not drive the result. The subject bootstrap also treats each subject's cross-validated score as fixed, and subjects sharing a fold share a fitted model; the leakage-free grouped bootstrap that refits the models inside each resample essentially reproduces the headline intervals and still excludes zero in all three cohorts (Inference robustness, Section 3.1), so this does not drive the result either. The transformer is scored on full context-length chunks, dropping each subject's final partial window while Markov-1 scores all tokens; this cancels in the paired contrast. The transformer is deliberately small and leaves a smaller residual than the classical higher-order Markov chains; a larger model might extract more, though the methodological point about the surrogate is model-agnostic. Finally, for very short or highly constrained sequences the bigram multigraph admits few Eulerian trails, so the surrogate ensemble has limited variability and the test loses power; our sequences are long (hundreds to thousands of tokens), so this is not a concern here, but it bounds the method for short streams. We replicated the corrected test across three independent adult cohorts (Table 4) and confirmed robustness to the duration-bucket edges; replication in clinical or paediatric samples remains to be done.

5 Conclusion

EEG microstate sequences carry a small but reliable amount of structure beyond a first-order Markov chain. The right way to measure it is a bigram-preserving surrogate, which keeps the bigram statistics the Markov-1 baseline uses and so isolates, from per-subject context, residual structure not explained by the preserved bigram counts. We recommend the bigram-preserving surrogate as a default conditional null for microstate sequence-model comparisons against a first-order baseline. Once it is applied, the residual that remains is small, can be captured by classical higher-order Markov models rather than only by a transformer, and, on the present data, has no detectable behavioural correlate.

Statements and Declarations

Funding. The author received no specific funding for this work.

Competing Interests. The author has no competing interests to declare that are relevant to the content of this article.

Ethics approval. This study is a secondary analysis of openly released, fully de-identified public datasets (LEMON; Dortmund Vital Study, OpenNeuro ds005385; SRM, OpenNeuro ds003775) and did not involve new data collection. Each dataset was collected and shared by its original providers under their respective ethics approvals and public data-use terms; the original LEMON acquisition was approved by the ethics committee of the University of Leipzig.

Consent to participate. Informed consent was obtained from participants by the original data providers; this study analysed only the de-identified public releases.

Consent for publication. Not applicable; no identifiable individual-level data are reported.

Data availability. This study is a secondary analysis of three openly available public EEG datasets; no new data were generated. LEMON (Babayan et al. 2019) is available from the FCP-INDI / MPI-Leipzig repository (http://fcon_1000.projects.nitrc.org/indi/retro/MPI_LEMON.html); the Dortmund Vital Study (Getzmann et al. 2024) is available from OpenNeuro under accession ds005385 (https://openneuro.org/datasets/ds005385); and the SRM resting-state set (Hatlestad-Hall et al. 2022) is available from OpenNeuro under accession ds003775 (https://openneuro.org/datasets/ds003775). Each dataset is distributed under its own open data-use terms.

Code availability. All code, configurations, per-subject negative log-likelihoods, and per-fold subject-ID splits are available to the editors and reviewers on request during peer review. On acceptance the full repository will be released publicly under the MIT licence with a Zenodo-archived DOI.

Author contributions. M.B.I. conceived the study, implemented the methods and software, performed the analysis, and wrote the manuscript.

Use of generative AI. During this work the author used AI-based coding and writing assistants to help implement analysis code and to draft and edit the manuscript. The author verified all analyses, figures and numerical results and takes full responsibility for the content; no AI system is listed as an author.

References

Stephen F. Altschul and Bruce W. Erickson. Significance of nucleotide sequence alignments: a method for random sequence permutation that preserves dinucleotide and codon usage. *Molecular Biology and Evolution*, 2(6):526–538, 1985. URL https://doi.org/10.1093/oxfordjournals.molbev.a040370.

Anahit Babayan, Miray Erbey, Deniz Kumral, Janis D. Reinelt, Andrea M.F. Reiter, Josefin Röbbig, 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. URL https://doi.org/10.1038/sdata.2018.308.

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

Markus M. Breunig, Hans-Peter Kriegel, Raymond T. Ng, and Jörg Sander. LOF: identifying density-based local outliers. In *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data*, pages 93–104, 2000. URL https://doi.org/10.1145/342009.335388.

Geoffrey Brookshire, Jake Kasper, Nicholas M. Blauch, Yunan Charles Wu, Ryan Glatt, David A. Merrill, Spencer Gerrol, Keith J. Yoder, Colin Quirk, and Ché Lucero. Data leakage in deep learning studies of translational EEG. *Frontiers in Neuroscience*, 18:1373515, 2024. URL https://doi.org/10.3389/fnins.2024.1373515.

Gustavo Deco, Viktor K. Jirsa, and Anthony R. McIntosh. Emerging concepts for the dynamical organization of resting-state activity in the brain. *Nature Reviews Neuroscience*, 12(1):43–56, 2011. URL https://doi.org/10.1038/nrn2961.

Victor Férat, Mathieu Scheltienne, Denis Brunet, Tomas Ros, and Christoph Michel. Pycrostates: a Python library to study EEG microstates. *Journal of Open Source Software*, 7(78):4564, 2022. URL https://doi.org/10.21105/joss.04564.

Stephan Getzmann, Patrick D. Gajewski, Daniel Schneider, and Edmund Wascher. Resting-state EEG data before and after cognitive activity across the adult lifespan and a 5-year follow-up. *Scientific Data*, 11:988, 2024. URL https://doi.org/10.1038/s41597-024-03797-w.

Alexandre Gramfort, Martin Luessi, Eric Larson, Denis A. Engemann, Daniel 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. URL https://doi.org/10.3389/fnins.2013.00267.

Markus Gschwind, Christoph M. Michel, and Dimitri Van De Ville. Long-range dependencies make the difference: comment on "a stochastic model for EEG microstate sequence analysis". *NeuroImage*, 117:449–455, 2015. URL https://doi.org/10.1016/j.neuroimage.2015.05.062.

Christoffer Hatlestad-Hall, Trine Waage Rygvold, and Stein Andersson. BIDS-structured resting-state electroencephalography (EEG) data extracted from an experimental paradigm. *Data in Brief*, 45: 108647, 2022. URL https://doi.org/10.1016/j.dib.2022.108647.

Daniel Haydock, Shabnam Kadir, Robert Leech, Chrystopher L. Nehaniv, and Elena Antonova. EEG microstate syntax analysis: a review of methodological challenges and advances. *NeuroImage*, 309: 121090, 2025. URL https://doi.org/10.1016/j.neuroimage.2025.121090.

Carl Hierholzer. Ueber die Möglichkeit, einen Linienzug ohne Wiederholung und ohne Unterbrechung zu umfahren. *Mathematische Annalen*, 6(1):30–32, 1873. URL https://doi.org/10.1007/BF01442866.

Fatemeh Zareayan Jahromy. Complexity and non-predictability in neurodynamic: gender-specific EEG dynamics uncovered via hidden Markov models. *Cognitive Neurodynamics*, 19(1):87, 2025. URL https://doi.org/10.1007/s11571-025-10271-9.

Minghui Jiang, James Anderson, Joel Gillespie, and Martin Mayne. ushuffle: a useful tool for shuffling biological sequences while preserving the k-let counts. *BMC Bioinformatics*, 9:192, 2008. URL https://doi.org/10.1186/1471-2105-9-192.

J. A. Scott Kelso. Multistability and metastability: understanding dynamic coordination in the brain. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 367(1591):906–918, 2012. URL https://doi.org/10.1098/rstb.2011.0351.

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

Reinhard Kneser and Hermann Ney. Improved backing-off for M-gram language modeling. *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 1:181–184, 1995. URL https://doi.org/10.1109/ICASSP.1995.479394.

Adam Li, Jacob Feitelberg, Anand Prakash Saini, Richard Höchenberger, and Mathieu Scheltienne. MNE-ICALabel: automatically annotating ICA components with ICLabel in Python. *Journal of Open Source Software*, 7(76):4484, 2022. URL https://doi.org/10.21105/joss.04484.

Sergei Maslov and Kim Sneppen. Specificity and stability in topology of protein networks. *Science*, 296 (5569):910–913, 2002. URL https://doi.org/10.1126/science.1065103.

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. URL https://doi.org/10.1016/j.neuroimage.2017.11.062.

Chrystopher L. Nehaniv and Elena Antonova. Simulating and reconstructing neurodynamics with epsilon-automata applied to electroencephalography (EEG) microstate sequences. In *2017 IEEE Symposium Series on Computational Intelligence (SSCI)*, pages 1753–1761, 2017. URL https://doi.org/10.1109/SSCI.2017.8285438.

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. URL https://doi.org/10.1109/10.391164.

Adam Paszke, Sam Gross, Francisco Massa, et al. PyTorch: an imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32 (NeurIPS 2019)*, pages 8024–8035, 2019. URL https://arxiv.org/abs/1912.01703.

F. Perrin, J. Pernier, O. Bertrand, and J. F. Echallier. Spherical splines for scalp potential and current density mapping. *Electroencephalography and Clinical Neurophysiology*, 72(2):184–187, 1989. URL https://doi.org/10.1016/0013-4694(89)90180-6.

Luca Pion-Tonachini, Ken Kreutz-Delgado, and Scott Makeig. ICLabel: an automated electroencephalographic independent component classifier, dataset, and website. *NeuroImage*, 198:181–197, 2019. URL https://doi.org/10.1016/j.neuroimage.2019.05.026.

Mikhail I. Rabinovich, Ramón Huerta, Pablo Varona, and Valentin S. Afraimovich. Transient cognitive dynamics, metastability, and decision making. *PLoS Computational Biology*, 4(5):e1000072, 2008. URL https://doi.org/10.1371/journal.pcbi.1000072.

Cosma Rohilla Shalizi and James P. Crutchfield. Computational mechanics: Pattern and prediction, structure and simplicity. *Journal of Statistical Physics*, 104(3–4):817–879, 2001. URL https://doi.org/10.1023/A:1010388907793.

Apoorva Sikka, Hamidreza Jamalabadi, Marina Krylova, Sarah Alizadeh, Johan N. van der Meer, Lena Danyeli, Matthias Deliano, Petya Vicheva, Tim Hahn, Thomas Koenig, Deepti R. Bathula, and Martin Walter. Investigating the temporal dynamics of electroencephalogram (EEG) microstates using recurrent neural networks. *Human Brain Mapping*, 41(9):2334–2346, 2020. URL https://doi.org/10.1002/hbm.24949.

Povilas Tarailis, Fiorenzo Artoni, Thomas Koenig, Christoph M. Michel, and Inga Griskova-Bulanova. Short-term and long-term test-retest reliability of memory, complexity, and randomness of EEG microstates sequence. *Cognitive Neurodynamics*, 20(1):19, 2025. URL https://doi.org/10.1007/s11571-025-10391-2.

Emmanuelle Tognoli and J. A. Scott Kelso. The metastable brain. *Neuron*, 81(1):35–48, 2014. URL https://doi.org/10.1016/j.neuron.2013.12.022.

Dimitri Van de Ville, Juliane Britz, and Christoph M. Michel. EEG microstate sequences in healthy humans at rest reveal scale-free dynamics. *Proceedings of the National Academy of Sciences*, 107 (42):18179–18184, 2010. URL https://doi.org/10.1073/pnas.1007841107.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems 30 (NIPS 2017)*, 2017. URL https://arxiv.org/abs/1706.03762.

Diego Vidaurre, Stephen M. Smith, and Mark W. Woolrich. Brain network dynamics are hierarchically organized in time. *Proceedings of the National Academy of Sciences*, 114(48):12827–12832, 2017. URL https://doi.org/10.1073/pnas.1705120114.

Frederic von Wegner and Helmut Laufs. Information-theoretical analysis of EEG microstate sequences in Python. *Frontiers in Neuroinformatics*, 12:30, 2018. URL https://doi.org/10.3389/fninf.2018.00030.

Frederic von Wegner, Enzo Tagliazucchi, and Helmut Laufs. Information-theoretical analysis of resting state EEG microstate sequences: non-Markovianity, non-stationarity and periodicities. *NeuroImage*, 158:99–111, 2017. URL https://doi.org/10.1016/j.neuroimage.2017.06.062.

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. URL https://doi.org/10.1016/j.compbiomed.2025.110367.

Frederic von Wegner, Gesine Hermann, Inken Tödt, Inga Karin Todtenhaupt, and Helmut Laufs. A quantitative comparison of two methods for higher-order EEG microstate syntax analysis. *Brain Topography*, 39(3):45, 2026. URL https://doi.org/10.1007/s10548-026-01196-5.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryCohortTemplates.pdf