Working notes: from “biomarker pipeline” to a dynamical microscope

Project slugs: - **phd-eeg-mci-biomarkers** (thesis + writeup + validation) - **eeg-state** (implementation + experiments + tooling)

This doc consolidates what we learned in the last sprint (autoencoder → latent trajectory → recurrence/RQA → optional state conditioning) and turns it into a forward plan, with an explicit **systems neuroscience** framing (Rabinovich-style metastable flows), while keeping **biomarker** as a secondary / “nice if it works” outcome.

---

1) What we built

1.1 Pipeline (current)

1) 256-channel EEG → Hilbert transform → per-channel instantaneous **phase** and **amplitude** 2) Input representation (per time sample): - **(cos φ, sin φ, log-amplitude)** per channel - 5s chunks @ 250 Hz → 1250 samples - Stacked channels → shape **(768, 1250)** 3) Transformer autoencoder compresses to latent trajectory: - latent trajectory: **(T=1250, D=64)** 4) From latent trajectory → recurrence matrix (RR-controlled) → **RQA features** 5) Classification (baseline): subject-level prediction using aggregated chunk RQA features 6) Optional “state discovery” and **state-conditioned** prediction with retention analysis + null controls

1.2 Model (as implemented)

TransformerAutoencoder (~427k params) - Encoder: conv stack 768→32→64→128 (BN+GELU) → proj 128→64 → TransformerEncoder (2 layers, 4 heads, FF=256) - Decoder: TransformerEncoder → proj 64→128 → ConvTranspose 128→64→32→768 - Output activations: tanh for cos/sin; linear for amplitude - Loss: - reconstruction MSE (phase + amplitude) + unit-circle regularizer - contrastive (SupCon-style, $\lambda \approx 0.1$, temp=0.07)

---

2) Key empirical outcomes (numbers we should not lose)

2.1 Training (amplitude+contrastive) looks healthy

- Best val_loss around **0.612** (epoch ~199)
- Contrastive term falls to near-zero
- No NaNs, stable curves

Interpretation: optimization is working; the model is learning a stable mapping. Whether it learns *clinically relevant* structure is a separate question.

2.2 Integration experiment (full report you pasted)

Latent mode: **RAW** - **Baseline AUC: 0.627 ± 0.119** (CI: [0.349, 0.869]) - **State-conditioned AUC: 0.622 ± 0.114** (CI: [0.346, 0.862]) - Effect (state – baseline): **−0.005**

Retention: - overall **96.72%** (MCI 96.66%, HC 96.84%) - states discovered ~23.9, used ~13.5

Bias control: - **State-conditioned AUC == retention-matched baseline AUC (Δ=0.000)** → selection/retention is explaining the effect, not "states" adding signal.

Trained vs random: - baseline: trained 0.629 vs random 0.623 (Δ +0.006) - state-conditioned: trained 0.617 vs random 0.626 (Δ −0.009)

Null sanity: - real state-conditioned 0.622 vs null (shuffled states) 0.617 (very close)

Takeaway: - **Baseline improved a lot** vs earlier phase-only (~0.44), which is a real step forward. - **State-conditioning is not providing added value** once retention matching is applied. - **Trained vs random is still unconvincing** at the level of this experiment.

2.3 Local quick tests: why they look "better" (but are fragile)

Your local runs can show AUC ~0.70+ because: - tiny test set (e.g., 4 subjects), high variance - train/test split differs each run - segment-level vs subject-level metrics can diverge

We should treat local AUCs as **sanity checks**, not evidence.

---

3) What "state conditioning" was supposed to mean (and why it currently doesn't help)

3.1 Original idea

1) Discover latent "states" (clusters / neighborhoods) in *train* latent space. 2) For each subject, keep only chunks that can be confidently assigned to a state (or a subset of "good" states). 3) Compute RQA features / classify **on retained chunks**.

Hypothesis: MCI vs HC differences might be **state-dependent** (only visible in certain dynamical regimes). State-conditioning was meant to: - reduce mixing of incompatible regimes - focus analysis on stable metastable regimes

3.2 What the retention-matched baseline reveals

If "state-conditioned" == "retention-matched baseline", then: - the benefit is coming from **throwing away chunks**, not from meaningful state structure.

In your latest run retention is ~97%, so selection is mild, but the equality still says: *state labels themselves are not adding predictive structure beyond the fact of filtering.*

3.3 What to do with this

- Keep state-conditioning as an **analysis lens** (describe regimes) rather than expecting it to boost AUC.
- If we want state labels to matter, we need either:
- a better state model (e.g., HMM on latent trajectory), or
- a training objective that explicitly organizes latent space into meaningful regimes.

---

4) Systems neuroscience pivot: the pipeline as a “dynamical microscope”

4.1 The core reframing

Instead of claiming “a clinical biomarker”, claim: - **a method to compress multichannel EEG into a low-dimensional trajectory that preserves dynamical structure**, - then quantify structure via recurrence + metastability metrics, - and use it to study **how brain dynamics differ across conditions**.

Clinical prediction becomes a secondary validation (“it correlates with diagnosis / cognition”), not the central novelty.

4.2 Rabinovich-style interpretation hooks

Rabinovich’s framework (as you summarized) emphasizes: - metastable states / sequences of transiently stable regimes - “information flow” as trajectories through phase space - disturbances in flow / stability as meaningful, not only discrete states

Your new trajectory plots (flow fields, dwell density, speed/metastability overlays) are directly compatible with that framing: - **Flow field**: local displacement vectors in latent space - **Dwell density**: where the system spends time (candidate metastable regions) - **Metastability via speed**: slow regions as “sticky” regimes

The HC vs AD example you showed (density becoming more “lumpy” / less clean circulation) is exactly the kind of *systems* story we can formalize.

4.3 What we can *legitimately* claim (when done rigorously)

With the right controls (below), we can claim: - the learned latent dynamics show **structured recurrence properties** beyond trivial autocorrelation - group differences exist in: - dwell-time distributions - transition structure (Markov/HMM) - recurrence network geometry - metastability (speed distribution / residence time) - flow-field stability (local divergence / curl proxies)

Biomarker claim requires extra steps (generalization, calibration, external validation). The dynamical microscope claim is more plausible and novel.

5) Theoretical + methodological next steps (what to implement next)

5.1 Replace “state discovery” with explicit dynamical models

You asked about Markov/HMM — yes, this is a natural next layer.

Option A — HMM on latent trajectory - Fit an HMM (Gaussian emissions) on latent trajectories per subject or pooled with subject random effects. - Outputs: - state occupancy (fraction time) - dwell time distribution per state - transition matrix - entropy rate / metastability indices - Compare these between groups (HC vs MCI vs AD).

Option B — Markov chain on discretized latent space - Discretize latent space via k-means / GMM / Voronoi bins. - Build transition matrix and compute: - stationary distribution - mixing time - metastable sets (PCCA+/spectral clustering)

Option C — Continuous-time / flow view - Estimate local flow field $F(x) \approx E[x_{t+\Delta} - x_t \mid x_t \text{ in bin}]$ - Derive summary metrics: - curl / divergence surrogates - stability of fixed points / attractor basins (empirical)

5.2 “Recurrence graph” analysis (bridges to network neuroscience)

- Treat recurrence matrix as adjacency → recurrence network
- Compute network measures:
- clustering coefficient
- path length
- modularity
- motif counts These have interpretable links to dynamical complexity and can be compared across groups.

5.3 Controls we *must* include to be credible

1) **Trained vs random weights** (but done right) - report distribution across multiple seeds - same downstream pipeline - show effect sizes, not single-run deltas

2) **Surrogate data** - within-chunk time shuffle (destroys temporal structure) - phase randomization / Fourier surrogate (preserve spectrum, destroy phase relations) - amplitude shuffle controls

3) **Leakage checks** - subject-wise splits (always) - any dimensionality reduction (PCA/UMAP) fit on train only

4) **Stability / reproducibility** - split-half within subject: do we recover similar flow/dwell maps? - day/session effects

5.4 Why trained≈random can happen (and how to address it)

- Untrained conv/transformer blocks impose smoothness/low-dimensional structure by architecture alone.
- If downstream summary features (RQA) mostly measure smoothness, training adds little.

How to make training matter: - increase reliance on learned structure by changing objective: - self-supervised predictive loss (next-step prediction) - masked reconstruction - subject-contrastive (positives = same subject across time) - multitask with diagnosis head (weak supervision) - evaluate not only AUC but **dynamical metrics stability** across time.

---

6) Updated 3-month execution plan (v2)

Month 1 — Lock correctness + build dynamical microscope "core analyses"

Goal: a rigorous, reproducible analysis package + first group-level dynamical results.

1) **Correctness + rigor pass (1–2 weeks)** - Fix/understand why RESIDUALIZED latent mode reports NaN. - Re-run integration experiment with: - multiple seeds - consistent folds (StratifiedGroupKFold) - probability-based AUC - retention-matched baseline + null model - Produce a "Methods sanity appendix" figure set: - trained vs random distributions - surrogate tests - leakage-free pipeline diagram

2) **Trajectory + metastability battery (1–2 weeks)** Implement per-subject metrics (computed from latent trajectories): - speed distribution (mean/std/CV; heavy tails) - dwell regions (density peaks) + dwell time distribution - displacement, path length, tortuosity (with scale normalization) - recurrence-network measures

Deliverables: - a single notebook/script that outputs a standardized **per-subject dynamical report**.

Month 2 — State-space modeling (HMM/Markov) + group comparisons

Goal: move from "pretty plots" to **quantified metastable organization**.

1) **HMM on latent trajectories** - choose model order via BIC/AIC + held-out likelihood - compare HC/MCI/AD in: - occupancy - dwell times - transition entropy - metastable sets

2) **Markov / spectral metastability (optional, if HMM too heavy)** - discretize latent space - compute metastable sets + mixing times

Deliverables: - 1–2 main figures + supplement: transition matrices, dwell histograms, group effect sizes.

Month 3 — Thesis story + writeup + "biomarker as secondary"

Goal: a defendable thesis narrative + a paper-style structure.

1) **Thesis framing** - Contribution #1: representation learning for multichannel EEG → low-D dynamical trajectory - Contribution #2: recurrence + metastability toolbox (dynamic microscope) - Contribution #3: clinical relevance as *validation*, not sole claim

2) **Write results** - group differences in dynamical metrics - robustness controls - modest classification results (if present) framed carefully

3) **Packaging** - clean CLI commands to reproduce analyses - deterministic configs

---

7) "What to tell Claude to implement next" (concrete backlog)

A) Minimal next implementation (1–3 days)

- Add HMM fitting on latent trajectories (per subject + pooled)
- Output: transition matrix + dwell distribution + occupancy
- Save plots + a CSV summary per subject

B) Medium (1–2 weeks)

- Surrogate generation (phase randomization, Fourier surrogate)
- Flow-field stability metrics (local divergence/curl proxies)
- Recurrence network analysis

C) Stretch

- Cross-subject latent alignment (Procrustes / CCA) to compare phase spaces meaningfully
- Conditional flow fields (task segments / sleep stage / eyes open vs closed)

---

8) Decision checkpoint (how we know we're "there")

We're ready to claim "dynamic microscope" when we can show: 1) **reproducible** within-subject dynamical signatures (split-half stability) 2) **group-level** differences with effect sizes + confidence intervals 3) robust to **surrogates** and **leakage** 4) interpretable links between metrics and known physiology (e.g., slowing, reduced flexibility, altered dwell structure)

We're ready to claim "biomarker" only if additionally: - performance is stable across folds/seeds - trained clearly beats random - validated on held-out cohort or external dataset

---

Appendix: short interpretation of your flow/density plots

- The "ring-like" trajectories often arise when the dominant components encode oscillatory modes; this is not automatically bad.
- The useful question is whether groups differ in:

- where the ring breaks / density lumps
- speed slowing / sticking events
- transition structure between dense regions
- Your HC vs AD example already suggests a hypothesis: **AD shows less uniform circulation and more residence in particular regions** (to be tested across subjects, not visually).