SEPARATING DIAGNOSIS FROM DISEASE REPRESENTATION: DUAL-VIEW EEG LEARNING WITH NEURAL-DYNAMICS-GUIDED DEFORMATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Electroencephalography (EEG)-based closed-loop neuromodulation calls for a subject-specific structured state, as opposed to a single disease probability, specifying which brain regions are deviant, at which frequencies, and at which lags. Sensor-space models keep the strongest diagnostic evidence without anatomy, source-space models give anatomy at a loss of predictive signal, and post-hoc attributions stay outside the prediction. We separate the two instead of forcing them into one representation, and propose DMD-EEG (Dual-view Multiscale Deformation for EEG), which keeps a fixed scalp spectral expert for diagnosis and models the source-space disease-related representation as a low-rank, sparse, iterative deformation of a healthy neural-dynamics prior in a 46-region-of-interest (ROI) $\times$ 5-frequency $\times$ 4-lag (autocorrelation-timescale) space. The two experts meet only at a fixed decision level, so the source state is architecturally separate from the scalp expert. Across major depressive disorder (MDD), first-episode psychosis (FEP), and Parkinson's disease (PD), decision-level fusion matches the strongest single expert on MDD and FEP and exceeds the source branch on PD. On FEP the source expert is the strongest branch, the task where the deformation contributes most. The source state is an explicit ROI–frequency–lag attribution defined in a shared source coordinate system across montages, which we treat as an anatomically-coordinated predictive representation whose coordinates are directly readable and hypothesis-generating. The highest-saliency coordinates align with established disease circuitry (fronto-limbic-temporal regions in MDD, motor-cortex beta in PD), and the MDD state transfers by rank to an unseen cohort recorded with a different montage.

1 INTRODUCTION

Diagnostic fidelity and representational readability are different quantities. Source reconstruction dilutes scalp spectral discrimination, by 0.08–0.17 area under the ROC curve (AUC) under the strongest power spectral density (PSD) classifier per disease (Welch, 1967) and by up to 0.16 AUC under a higher-capacity EEGNet (Lawhern et al., 2018) (Appendix D), yet only the source space carries anatomical meaning. A source-only bottleneck therefore trades away diagnostic signal, and feature concatenation contaminates the source state with sensor-level discriminant information. The objective that maximizes accuracy differs from the objective that yields a readable, anatomically-coordinated disease representation. Clinical and neuromodulation use of EEG needs a structured, real-time state beyond a disease label: where a deviant state occurs, at what frequency, and with what temporal relationship. The central problem is a stable coordinate system for subject-specific neural deviation.

Existing work falls into three groups, none closing the gap. Sensor/source fusion studies fuse at the feature level (Shim et al., 2016; Huang et al., 2026). Without explicit separation constraints, a fused representation leaves the separate contributions of its sensor and source components unidentified (Park et al., 2025; Locatello et al., 2019). Source-space connectivity studies provide ROI-, frequency-, and lagged connectivity, but through fixed estimators or static features (Damborská et al.,

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.

2020; Hasanzadeh et al., 2020; Liu et al., 2022; Yang et al., 2024; Chen et al., 2024; Myers et al., 2025; Whitton et al., 2018), describing a population-level summary. Post-hoc feature-attribution and visualization methods can reveal a trained predictor's reliance on input features, but discriminative weights in general fall short of identifying the neural sources or mechanisms generating those features (Haufe et al., 2014; Schirrmeister et al., 2017). No existing method preserves strong sensor diagnosis, keeps a source representation independent, and makes the disease-related signal an explicit coordinate of the prediction itself.

We address these requirements with two experts combined at a low-capacity decision level: a fixed scalp expert mapping power-spectral evidence to a diagnostic probability, and a source-space expert expressing the disease-related state as an iterative sparse deformation of a subject-conditioned healthy-anchored state h^0. The branches share no intermediate learned features or training gradients and interact only at the decision level. With the deformation defined on explicit ROI, frequency, and lag coordinates, the source expert localizes the deviation in terms that can be checked against independent medical literature. We evaluate on three clinical tasks spanning psychiatric and neurological disorders – major depressive disorder (MDD, MODMA, 24 MDD/29 healthy controls (HC) (Cai et al., 2022)), first-episode psychosis (FEP, ds003944, 50 FEP/32 HC (Salisbury et al., 2022a)), and Parkinson's disease (PD, ds004584, 100 PD/49 HC (Singh et al., 2024)) – under a leak-free multi-seed subject-level holdout. These datasets lack subject-level work under a unified protocol. We supply that cross-disease comparison as the benchmark.

Our contributions are:

- **Dual-view separation of diagnosis and disease representation.** We separate diagnostic fidelity from representational readability in clinical EEG: a fixed low-capacity scalp expert handles diagnosis while a source-space expert learns a disease-related representation in anatomically labeled coordinates, combined only at a fixed decision level. The fusion matches the best single expert within seed noise on MDD and FEP (MDD 0.700 vs. 0.622; FEP 0.789 vs. 0.794), halves the source branch's across-seed standard deviation on FEP ($0.139 \rightarrow 0.055$), and exceeds the source branch on PD ($+0.176$ paired ΔAUC).
- **Neural-dynamics-guided iterative deformation.** The classification signal is carried by a low-rank iterative sparse deformation of the healthy-anchored state h^0 with explicit ROI $\times$ frequency $\times$ lag coordinates, so the decision is readable directly on anatomically labeled axes.
- **Structured, neurobiologically testable localization.** The decision-sensitivity profile over the healthy-anchored state carries explicit ROI $\times$ frequency $\times$ lag coordinates whose highest-saliency entries are directionally consistent with established disease circuitry, and the state is independently classifiable and, for MDD, rank-transferable across acquisition systems. We present this attribution as an anatomically-coordinated, hypothesis-generating predictive representation. It is supported by sensitivity attribution, ROI occlusion, and concordance with the disease literature: MDD on fronto-limbic-temporal regions, PD on motor-cortex beta.

2 Related Work

2.1 Sensor- and source-space EEG for brain disorder recognition

Sensor-level EEG features have been used for disorder recognition across MDD, FEP, and PD, with recordings, feature definitions, and evaluation protocols varying substantially (Sun et al., 2020; Redwan et al., 2024; Jaramillo-Jimenez et al., 2023). Deep and foundation models on scalp EEG form a parallel representation-learning line (Schirrmeister et al., 2017; Jiang et al., 2024; Wang et al., 2025). Source-space studies add anatomy through reconstruction and connectivity, often with fixed estimators (Damborská et al., 2020; Hasanzadeh et al., 2020; Liu et al., 2022; Yang et al., 2024; Chen et al., 2024). Feature-level sensor–source fusion learns a joint representation, but without explicit separation constraints the contribution of each representation stays ambiguous (Park et al., 2025; Locatello et al., 2019). Consequently, a feature-level fusion entangles the source component with the sensor features, leaving the source representation hard to identify on its own. Explicit bottleneck variables keep an intermediate representation that can be inspected and intervened on directly (Koh et al., 2020). Our source state is a learned deformation in anatomical coordinates, distinct from their human-defined concepts. Prior work asks whether combining representations improves classi-

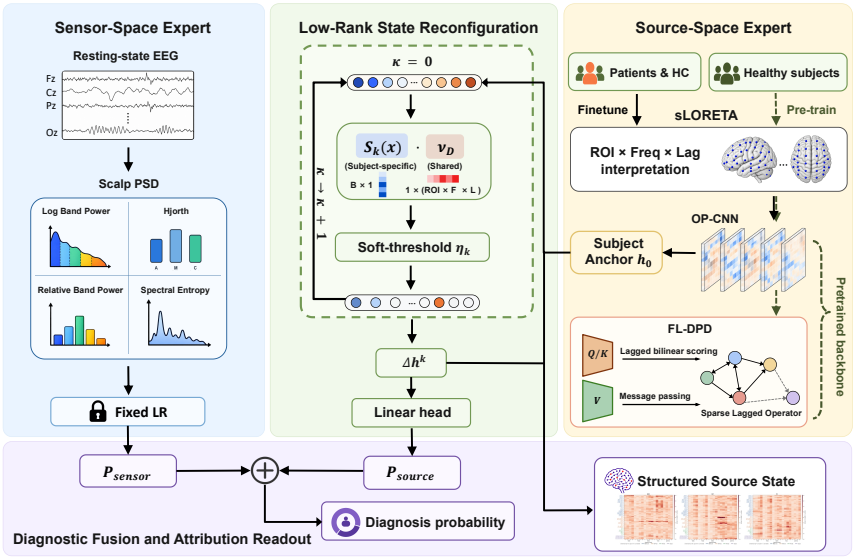


Figure 1: **Dual-view EEG learning.** A fixed PSD–logistic-regression sensor expert maps scalp spectra to a diagnostic probability, and a source expert anchors sLORETA activity to a healthy-dynamics prior (pretrained on healthy EEG only) and deforms it with a low-rank sparse solver into an explicit ROI×frequency×lag state. The two experts combine only at a fixed decision level, leaving the two branches architecturally separate and the source state available for attribution.

fication. We ask how to combine predictive evidence while preserving a source representation for attribution analysis.

2.2 Directed and Frequency-Specific Brain Communication

Directed and frequency-specific connectivity is well established (Granger causality (Granger, 1969), directed transfer function (Kamiński & Błinowska, 1991), partial directed coherence (Baccalá & Sameshima, 2001), phase transfer entropy (Lobier et al., 2014)) and applied to depression and psychosis EEG (Damborská et al., 2020; Hasanzadeh et al., 2020; Liu et al., 2022; Yang et al., 2024; Chen et al., 2024; Kabbara et al., 2022). Intracranial electrophysiology has likewise been used to examine directed prefrontal connectivity in MDD (Myers et al., 2025). These estimate a graph for classification or group comparison, whereas our source disease representation is a learned deformation state under a healthy-dynamics prior, retained for downstream analysis.

2.3 Counterfactual, Interpretable, and Neurobiologically Validated Modeling

Post-hoc feature attribution and visualization provide useful model interpretation, but by themselves leave a forward disease representation undefined (Haufe et al., 2014; Schirrmeister et al., 2017). Normative modeling frames patient deviation from a healthy distribution (Marquand et al., 2016; 2019), and data-driven subtype models decompose heterogeneous disorders into trajectories (Chen et al., 2023; Young et al., 2018). Source-space EEG normative models predict clinical outcomes from each subject's deviation from a healthy norm (Tong et al., 2024). Our deformation is supervised directly by disease labels, which makes it a discriminative state, distinct from an estimated healthy-counterfactual deviation. It carries the classification decision and is defined on explicit neurophysiological coordinates. Each coordinate of the healthy-anchored state has one neural meaning, so it is testable on two grounds: state stability and coordinate readability, and concordance of the highest-saliency coordinates with known disease circuitry. ROI, network, and frequency findings are checked against disease neuroscience, and lag through cross-fold stability, synthetic recovery, and lag perturbation (Appendix I.7).

3 METHOD

3.1 OVERVIEW

The framework combines two independent experts through a low-capacity fusion (Figure 1). The Sensor Expert maps scalp PSD (Welch, 1967) through a fixed logistic regression to p_{sensor}. The Source Expert source-reconstructs the EEG with a healthy-pretrained oscillatory model and applies a low-rank iterative deformation. A linear head then reads the resulting state to produce p_{source}. The final decision is the fixed average of the two experts,

$$p_{\text{final}} = \tfrac{1}{2}p_{\text{sensor}} + \tfrac{1}{2}p_{\text{source}}. \tag{1}$$

which keeps the two branches architecturally separate and the scalp expert purely diagnostic. A learnable weight α is evaluated as a secondary ablation (Appendix F.1).

3.2 ROI ATLAS DESIGN

We use a fixed 46-ROI analysis atlas, chosen before final model evaluation to balance transdiagnostic coverage against the resolution scalp-EEG source reconstruction can support. The ROIs are label-independent and cover MDD, FEP, and PD circuits, so all three disorders share one anatomical coordinate system. Deep structures such as basal ganglia are excluded because routine scalp EEG and sLORETA cannot resolve their subnuclei, so PD interpretation is restricted to cortical, thalamocortical, and large-scale network consequences. The full list, with the design rationale for each region, is in Table 5.

3.3 HEALTHY-DYNAMICS PRETRAINING

Healthy pretraining never sees a disease label. Scalp EEG is source-reconstructed with sLORETA (Pascual-Marqui, 2002) on a template head model, and each ROI's time series is then decomposed into five bands $(\delta, \theta, \alpha, \beta, \gamma)$. The oscillation-preserving CNN (OP-CNN) is a causal dilated-convolution encoder that preserves ROI, band, and temporal structure. The lag axis uses the per-ROI, per-band autocorrelation function (ACF) of the latent time series at $\tau \in \{1, 2, 4, 8\}$ (scale axis), reduced to one scalar per lag, yielding $h_H \in \mathbb{R}^{46\times5\times4}$ with genuine autocorrelation timescales, free of synthesized coordinates or inter-regional transmission delays. Pretraining optimizes local self-prediction plus cross-ROI communication reconstruction on healthy data only, jointly training the sparse static scaffold W_H and the frequency–lag directed pathway (FL-DPD) generator that produces a communication state

$$A_H = W_H + G_H, \qquad G_H = \tanh\big(h_{H,i} + h_{H,j} + b_{f,\ell}\big), \tag{2}$$

with G_H normalized per sample. Pretraining therefore learns a *healthy neural-dynamics prior*: source EEG maps to an oscillatory local state and then to cross-ROI delayed prediction. The healthy-only corpus pools 269 subjects from seven healthy datasets (Section 4.1).

3.4 NEURAL-DYNAMICS-CONDITIONED ITERATIVE DISEASE DEFORMATION

The healthy model provides each subject with a subject-conditioned anchor: the band-preserving source-space representation from the oscillation-preserving encoder, with a real per-τ autocorrelation lag axis, compressed by a small multilayer perceptron (MLP) applied independently per (ROI, band, lag) coordinate into $h^0 = h_H(x) \in \mathbb{R}^{920} = \mathbb{R}^{46\times5\times4}$, so each output coordinate keeps its ROI, band, and lag identity. Here "healthy" denotes the pretraining objective that shapes the encoder into a healthy-dynamics prior, distinct from a frozen template: each anchor comes from that encoder applied to the subject's own source activity. We write h^0 as h_{health} when we attribute the decision to its coordinates. Disease is then a K-step sparse deformation of h^0,

$$h^k = h^0 + \Delta h^k, \qquad \Delta h^0 = 0, \qquad k = 1, \ldots, K. \tag{3}$$

The deformation is restricted to a low-rank, cohort-shared disease subspace. A shared direction matrix $V_D \in \mathbb{R}^{r\times920}$ defines r disease axes, and a small amplitude network predicts a per-subject, per-step signed coefficient $s_k(x) \in \mathbb{R}^r$ from the anchor state $h^0(x)$, conditioning each step on the subject's neural dynamics,

$$U_k(x) = V_D^\top s_k(x), \tag{4}$$

so all subjects traverse the same shared disease directions with different signed distances, a strong capacity prior for small cohorts. This is an explicit modeling choice, fixing the subject-conditioned quantity as the signed position along a few disease axes shared across the cohort. We refer to this iterative deformation module as the solver. Each solver step applies three operations in sequence. First, the current deformation is moved along the low-rank direction with a per-step step size η_k:

$$\tilde{\Delta}h^{k+1} = \Delta h^k - \eta_k\, U_k(x). \tag{5}$$

Second, a per-step soft threshold τ_k shrinks small entries toward zero. Third, a hard top-m selection keeps only the m largest-magnitude coordinates, so at most m of the 920 entries are nonzero at each step:

$$\Delta h^{k+1} = \mathrm{TopK}_m\Big[\,\mathrm{SoftThreshold}\big(\tilde{\Delta}h^{k+1},\, \tau_k\big)\Big], \tag{6}$$

where $\mathrm{SoftThreshold}(z,\tau) = \mathrm{sign}(z)\max(|z|-\tau, 0)$. The budget m=460 (top-half of the 920 coordinates, fixed before training) caps the deformation at m locatable ROI–band–lag coordinates. A fixed fraction is preferable to a single magnitude threshold, which can collapse the support as per-subject magnitudes drift. Retaining the top half prunes the low-magnitude tail while keeping enough candidate coordinates, balancing representational richness against signal purity. Step size and threshold are learned per step and bounded by the solver's parameterization, $\eta_k \in [10^{-4}, 10^{-1}]$ and $\tau_k \in [0, 10^{-3}]$. Because each step adds $-\eta_k U_k$ to all 920 coordinates before selection, the top-m support is re-estimated at every step: a coordinate suppressed at step k can re-enter at step $k+1$ once its magnitude returns to the top m. Each step carries its own amplitude, step size, and threshold, so the K steps jointly refine one sparse deformation and Δh^K is a nonlinear function of the per-step signed amplitudes $s_1(x), \ldots, s_K(x)$. The decision is read by a light linear head applied directly to Δh^K, with the solver parameters trained together with the backbone. The deformation Δh^K is therefore a source-space discriminative representation learned under disease-label supervision. The full communication pathway $A^K = W_H + G(h^K)$ reconstructs cross-ROI healthy dynamics during pretraining but stays outside the decision readout, since re-fitting it on tens of disease subjects would overfit, and we retain it as a component for larger cohorts. The solver is trained with the minimal objective

$$\mathcal{L} = \mathrm{CE}\big(\hat{y}^K, y\big) + \lambda\ \mathrm{mean}\big(|\Delta h^K|\big), \tag{7}$$

with $\lambda = 0.01$ unless stated otherwise. The ℓ_1 term suppresses deformation entries that fail to earn their place, without forcing $\Delta h^K = 0$. The per-step update in Eq. 6 never uses the disease label, so inference on unseen subjects is label-free: $x_{\mathrm{unknown}} \to \Delta h^1 \to \cdots \to \Delta h^K \to \hat{y}$. The label is used in the outer objective to train the network, and the trajectory $(\Delta h^1, \ldots, \Delta h^K)$ is an output of the forward computation. Because the ROI $\times$ frequency $\times$ lag structure of Δh^K is a model output, each coordinate carries its own ROI and band semantics. The deformation is therefore comparable across subjects and disorders in one coordinate system. Three quantities answer different questions and are reported as distinct layers: Δh^K is the deformation that carries the classification signal, $\partial\,\mathrm{logit}/\partial h_{\mathrm{health}}$ is the decision-sensitivity profile on the anchor from which our attribution is computed, and the Cohen's d of h^K is a descriptive group-difference map.

3.5 Independent sensor diagnostic expert

The sensor expert is fixed and simple. For each scalp channel, a power spectral density estimate (Welch, 1967) on 1–45 Hz segments is summarized into absolute, relative, and log band power per canonical band $(\delta, \theta, \alpha, \beta, \gamma)$, plus spectral entropy and the Hjorth activity, mobility, and complexity descriptors (Hjorth, 1970), and mapped through a logistic regression to p_{sensor}. This preserves high-fidelity spectral diagnostic evidence without disease-specific classifier selection or contamination of the source latent. Strong cross-validated classifiers are reported separately as baselines.

4 Experiments

4.1 Experimental Setup

Datasets. The backbone is pretrained on 269 healthy subjects across seven datasets (Computational Neuroimaging and Neuroengineering Lab at the University of Oklahoma, 2025; Ding et al., 2022;

Table 1: Clinical cohorts. All model results use strict subject-level splits.

Disease	Dataset	Patients	HC	Elec.	Rate
MDD	MODMA (Cai et al., 2022)	24	29	128	250 Hz
FEP	ds003944 (Salisbury et al., 2022a)	50	32	60	500 Hz
PD	ds004584 (Singh et al., 2024)	100	49	64	500 Hz

Kim et al., 2025b;a; Melcón et al., 2025; Kinley & Becker, 2025; Ghaffari et al., 2025; Anonymous, 2025; Wu et al., 2026), with subject-level grouping enforced. These corpora are separate from the three disease datasets, and subject-ID disjointness from the MODMA, ds003944, and ds004584 HCs is verified. Disease adaptation is evaluated on the three tasks of Table 1. All signals are band-pass filtered (0.5–70 Hz) with a 50 Hz notch, re-referenced to the common average, and source-reconstructed over the same 46-ROI atlas.

Evaluation Protocol. All model results use a leak-free protocol: for each seed $s \in \{42, 99, 2026\}$, a disjoint 20% subject holdout is locked as the test set, the remaining pool is split into 3 subject-level folds used only to select the fold-best checkpoint, and the inner-best model is evaluated exactly once on that seed's holdout. Holdouts are non-overlapping and excluded from training, early stopping, and selection. The reported value is mean$\pm$std over n=3 seeds. We report AUC, balanced accuracy (BAcc), area under the precision-recall curve (AUPRC), F1, and cross-seed standard deviation. A fixed five-fold protocol is used only for development ablations (Section 4.4) and is noted as selection-biased.

Baselines. We reproduced the principal subject-level baselines under the same strict holdout protocol: scalp PSD (Welch, 1967) with logistic regression, linear support vector machine (SVM) (Cortes & Vapnik, 1995), and random forest (Breiman, 2001); EEGNet (Lawhern et al., 2018); DeeperBrain (Wang et al., 2026); CBraMod (official checkpoint); microstate+SVM for FEP (Hill et al., 2026); and spectral random forest for PD (Chen et al., 2026). Table 2 (Appendix C) lists all reproduction results with DMD-EEG-Source and DMD-EEG (Full). Scalp PSD + LR is also the fixed Sensor Expert inside DMD-EEG. Each learning-based baseline used its own preprocessing, scaling, segmentation, and, where available, official code and checkpoint. Published participant-level reports on these datasets (Wang et al., 2026; Lin et al., 2026; Hill et al., 2026; Chen et al., 2026; Del Pup et al., 2025) use different protocols, so their numbers resist direct comparison. Every method is reported under one unified subject-level protocol.

4.2 Main Results

Diagnostic fidelity and anatomical structure in the two views. With the same PSD classifier applied to both representations, sLORETA source reconstruction dilutes scalp spectral discrimination by 0.08 to 0.17 AUC (source-minus-scalp penalty -0.08 to -0.17; Appendix Table 4): the source space gives anatomical coordinates, but the source-transformation pipeline (inverse reconstruction, regularization, and ROI aggregation) discards part of the sensor-level spectral signal. A stronger classifier recovers only part of this loss: EEGNet (Lawhern et al., 2018) applied to both representations under the identical holdout brings the source representation within 0.02 AUC of its scalp counterpart on MDD and FEP and 0.16 AUC below on PD (Table 4). Table 3 reports the dual-view results: the source branch is the strongest single expert on FEP (AUC 0.794), and the fixed fusion matches the best single branch within seed noise on MDD and FEP and exceeds the source branch on PD (MDD 0.700, FEP 0.789, PD 0.780), with a significant gain over the source branch on PD and a positive trend on MDD, and a preserved strong FEP source branch at half its across-seed standard deviation ($0.139 \rightarrow 0.055$).

Sensor and source experts capture complementary evidence. The two views cover different errors, which is what the fusion exploits: their pooled out-of-sample predictions are weakly to moderately correlated (Spearman r=0.46 MDD, 0.25 FEP, 0.17 PD), so they are non-redundant. A paired subject-level bootstrap (Table 9) gives fusion-minus-source AUC of $+0.067$ MDD, $+0.030$ FEP, $+0.176$ PD, and fusion-minus-sensor of $+0.052$, $+0.065$, -0.006. The PD fusion-minus-source gain has a 95% CI that excludes zero. On MDD the gains are positive trends, and on FEP the

Table 2: Strict subject-level holdout comparison under a unified local reproduction protocol (three seeds; mean±std). All learning-based baselines were reproduced under the same subject-level evaluation protocol. Scalp PSD + LR is also used as the fixed Sensor Expert in DMD-EEG; DMD-EEG-Source is the source-space neural-dynamics deformation branch; DMD-EEG (Full) combines the two experts with fixed equal-weight decision fusion (α=0.5). AUPRC and $\mathrm{F1}_{\mathrm{pos}}$ use disease as the positive class. Within each task and metric, **bold** marks the best result and underline the second best.

Task	Method	BAcc	AUC	AUPRC	$\mathrm{F1}_{\mathrm{pos}}$
MDD	Scalp PSD + LR (Welch, 1967)	**0.656±0.106**	0.622±0.129	0.652±0.169	**0.598±0.109**
MDD	Scalp PSD + RF (Breiman, 2001)	0.594±0.153	0.667±0.109	0.716±0.125	0.515±0.205
MDD	Scalp PSD + SVM (Cortes & Vapnik, 1995)	0.478±0.028	0.344±0.042	0.463±0.048	0.395±0.088
MDD	EEGNet (Lawhern et al., 2018)	0.578±0.058	0.522±0.087	0.599±0.078	0.555±0.060
MDD	DeeperBrain (Wang et al., 2026)	0.500±0.000	0.589±0.016	0.608±0.044	0.417±0.295
MDD	CBraMod (Wang et al., 2025)	0.483±0.024	0.556±0.042	0.548±0.021	0.550±0.106
MDD	**DMD-EEG-Source**	0.622±0.110	0.622±0.134	0.663±0.140	0.548±0.034
MDD	**DMD-EEG (Full)**	0.628±0.136	**0.700±0.109**	**0.758±0.110**	0.583±0.118
FEP	Scalp PSD + LR (Welch, 1967)	0.656±0.077	0.700±0.062	0.824±0.034	0.723±0.069
FEP	Scalp PSD + RF (Breiman, 2001)	0.517±0.027	0.617±0.041	0.768±0.071	0.665±0.026
FEP	Scalp PSD + SVM (Cortes & Vapnik, 1995)	0.500±0.000	0.683±0.095	0.801±0.058	0.769±0.000
FEP	EEGNet (Lawhern et al., 2018)	0.611±0.046	0.639±0.083	0.753±0.047	0.761±0.064
FEP	DeeperBrain (Wang et al., 2026)	0.500±0.000	0.639±0.091	0.791±0.032	0.769±0.000
FEP	CBraMod (Wang et al., 2025)	0.511±0.052	0.406±0.102	0.621±0.066	0.699±0.082
FEP	Microstate+SVM (Hill et al., 2026)	0.500±0.000	0.417±0.130	0.625±0.079	0.769±0.000
FEP	**DMD-EEG-Source**	**0.706±0.131**	**0.794±0.139**	**0.894±0.081**	**0.791±0.031**
FEP	**DMD-EEG (Full)**	0.689±0.068	0.789±0.055	0.883±0.036	0.765±0.049
PD	Scalp PSD + LR (Welch, 1967)	**0.700±0.035**	0.772±0.014	0.871±0.022	0.784±0.035
PD	Scalp PSD + RF (Breiman, 2001)	0.642±0.094	0.718±0.058	0.850±0.025	0.796±0.064
PD	Scalp PSD + SVM (Cortes & Vapnik, 1995)	0.575±0.074	**0.803±0.051**	**0.878±0.028**	**0.809±0.024**
PD	EEGNet (Lawhern et al., 2018)	0.658±0.065	0.743±0.087	0.854±0.046	0.600±0.062
PD	DeeperBrain (Wang et al., 2026)	0.592±0.066	0.680±0.064	0.825±0.040	0.762±0.027
PD	CBraMod (Wang et al., 2025)	0.642±0.082	0.712±0.032	0.834±0.008	0.611±0.171
PD	Spectral RF (Chen et al., 2026)	0.542±0.031	0.702±0.104	0.840±0.067	0.774±0.007
PD	**DMD-EEG-Source**	0.608±0.077	0.638±0.020	0.767±0.032	0.783±0.013
PD	**DMD-EEG (Full)**	0.683±0.012	0.780±0.004	0.864±0.022	0.792±0.025

Table 3: **Complementarity of sensor and source views in DMD-EEG.** The Sensor Expert is a fixed Scalp PSD + logistic-regression classifier; the Source Expert predicts directly from the deformation state Δh (DMD-EEG-Source), with solver depth selected per dataset (K=4, r=1 for MDD; K=3, r=1 for FEP; K=2, r=2 for PD). DMD-EEG (Full) fuses the two experts with fixed equal weights (α=0.5). Sensor–Source Corr. is the Spearman correlation between the pooled out-of-sample predictions of the two experts; $\mathrm{F1}_{\mathrm{pos}}$ is the positive-class F1. Within each disease and metric, **bold** marks the best branch and underline the second best.

Disease	Branch	BAcc	AUC	AUPRC	$\mathrm{F1}_{\mathrm{pos}}$	Sensor–Source Corr.
MDD	Sensor Expert (Scalp PSD + LR)	**0.656 ± 0.106**	0.622 ± 0.129	0.652 ± 0.169	**0.598 ± 0.109**	
MDD	Source Expert	0.622 ± 0.110	0.622 ± 0.134	0.663 ± 0.140	0.548 ± 0.034	0.46
MDD	**DMD-EEG (Full)**	0.628 ± 0.136	**0.700 ± 0.109**	**0.758 ± 0.110**	0.583 ± 0.118	
FEP	Sensor Expert (Scalp PSD + LR)	0.656 ± 0.077	0.700 ± 0.062	0.824 ± 0.034	0.723 ± 0.069	
FEP	Source Expert	**0.706 ± 0.131**	**0.794 ± 0.139**	**0.894 ± 0.081**	**0.791 ± 0.031**	0.25
FEP	**DMD-EEG (Full)**	0.689 ± 0.068	0.789 ± 0.055	0.883 ± 0.036	0.765 ± 0.049	
PD	Sensor Expert (Scalp PSD + LR)	**0.700 ± 0.035**	0.772 ± 0.014	**0.871 ± 0.022**	0.784 ± 0.035	
PD	Source Expert	0.608 ± 0.077	0.638 ± 0.020	0.767 ± 0.032	0.783 ± 0.013	0.17
PD	**DMD-EEG (Full)**	0.683 ± 0.012	**0.780 ± 0.004**	0.864 ± 0.022	**0.792 ± 0.025**	

fusion matches the strong source branch while halving its across-seed standard deviation. Appendix ROC curves (Figure 5) and the confidence scatter (Figure 6) show the same pattern at the subject level. On PD the signal is largely spectral: the scalp classifier is the strongest of any task-and-view combination (AUC 0.803), both source probes fall well below it (source-ROI PSD 0.640, source EEGNet 0.587), consistent with cortical power structure that source reconstruction dilutes plus deep basal-ganglia pathology sLORETA cannot resolve. The source view is thus complementary there, and PD is where the fusion gain over the weaker branch is significant (ΔAUC +0.18).

The deformation is an independently predictive, direction-reproducible discriminative axis. The source state is independently classifiable (MDD AUC 0.622, FEP 0.794, PD 0.638; Table 3), and its readout geometry is reproducible: the linear readout direction is cross-seed stable for MDD (signed cosine 0.971, permutation $p<0.001$) and FEP (0.840, $p<0.001$) in $2/3$ seed pairs, so the model sorts subjects along one consistent axis and every coordinate of the healthy-anchored state maps the decision to a specific ROI, band, and lag. Because single-fold estimates are noisy at this cohort size, we report the consolidated saliency (coordinates direction-consistent across the 3 splits $\times$ 3 folds; Table 14, Figure 2). The literature prior is fixed in the appendix and never used for training, checkpoint selection, Top-K selection, or hyperparameter tuning.

4.3 Analysis

Attribution readout and medical concordance. We distinguish three quantities. The anchor h^0 is the subject-conditioned, input-defined state whose coordinates carry a stable per-subject anatomical identity, and it is the object the attribution is computed on. Δh^K is the solver's deformation carrying the classification signal, and the final state $h^K = h^0 + \Delta h^K$ defines the descriptive group-effect map. The attribution below is the decision-sensitivity profile $\partial \,\mathrm{logit}/\partial h_{\mathrm{health}}$, which identifies the anchored coordinates the prediction depends on. Sensitivity and group difference measure different properties: sensitivity measures how much the decision depends on a coordinate, while the group difference measures how much the state varies between groups. A decision-critical coordinate can show little group difference, and a group-different coordinate can leave the decision unchanged, so we report the two separately. High-sensitivity ROIs are hypothesis-generating and prioritize plausible regions for future stimulation-target validation. We read the decision sensitivity from the healthy-anchored state by the unsigned amplitude saliency $\mathrm{mean}_{\mathrm{subj}}|\partial \,\mathrm{logit}/\partial h_{\mathrm{health}}|$, which is cross-seed stable (axis cosine 0.83–0.87) and agrees with the occlusion test (ROI Spearman $+0.31$ MDD, $+0.59$ PD; Appendix I.9). The signed saliency supplies the band and lag direction (Table 14, Figures 2, 7). The highest-importance ROIs concentrate on disease-relevant circuitry. For MDD, the prefrontal, temporal, limbic, and default-mode coordinates (DLPFC, MTG, amygdala, parahippocampal cortex, dACC, sgACC, dmPFC, PCC) are broadly concordant with reported large-scale network dysconnectivity in depression (Kaiser et al., 2015; Damborská et al., 2020). Independent structural work also reports lower hippocampal volume in MDD (Sheline et al., 1996; Videbech & Ravnkilde, 2004; Schmaal et al., 2016). For PD, the sensorimotor coordinates (M1, premotor, S1, SMA) and beta-weighted profile are qualitatively concordant with reported abnormal motor-cortical beta–gamma coupling and cortico-subthalamic beta connectivity (de Hemptinne et al., 2013; Litvak et al., 2011). They fall short of establishing elevated beta power in every individual source ROI. For FEP, the fronto-parietal and insular coordinates are consistent with meta-analytic evidence of dysconnectivity across default-mode, salience, and central-executive networks in early psychosis (O'Neill et al., 2019). As an interventional check, zeroing a top-ranked ROI's band–lag block lowers the decision AUC for FEP (SPL, dmPFC, insula), PD (premotor, M1, dACC), and MDD ($dACC_R$; a joint ROI set lowers it further, Appendix I.9), so the decision relies on anatomically plausible coordinates (the atlas is fixed before evaluation, and the literature played no role in training or selection).

External generalization. As an external check the trained MDD source branch transfers, in the shared 46-ROI space, to an independent cohort with a different montage (ds003478 (Cavanagh, 2021), 86 subjects: 11 current-MDD, 75 low-BDI controls): the ranking AUC is 0.724 ± 0.042 with MDD scoring above controls across seeds, though a fixed 0.5 threshold misses the ordering, so the transfer evidence is confined to ranking (Table 11).

4.4 Ablation Study

The ablations are controlled single-variable diagnostics (Appendix F). The one-shot (K=1) and no-deformation (K=0) controls stay below the selected depth on every task (Table 7). The no-deformation control feeds the anchor h^0 to the same linear head the solver path uses, isolating the deformation. On FEP the selected solver (K=3) reaches holdout AUC 0.794 against 0.678 at K=0, and MDD inner-validation AUC drops from 0.801 to 0.780, so the deformation carries discriminative signal beyond the anchor it is built from. Removing the healthy pretraining degrades the source branch on every task, most sharply on PD (strict-holdout AUC $0.638 \rightarrow 0.415$; Table 8),

so the healthy-dynamics pretraining helps the source branch. Removing the multiscale temporal context degrades MDD ranking with BAcc nearly unchanged (Appendix F). Solver depth and rank are selected per dataset by inner validation (K=4, r=1 MDD; K=3, r=1 FEP; K=2, r=2 PD; Table 7).

5 Discussion

5.1 The value of separation

Diagnosis and disease representation have different requirements: a scalp spectral expert preserves the diagnostic evidence that source reconstruction dilutes (Table 4), while a source-space expert produces an anatomically interpretable state, and the two combine at a fixed decision level. The fusion matches the best single branch within seed noise on MDD and FEP, raising MDD AUC from 0.622 to 0.700 and preserving the strong FEP source branch at roughly half its variance. On PD it exceeds the source branch, from 0.638 to 0.780. Sensor–source correlation stays low as the source branch strengthens (pooled Spearman $0.28 \rightarrow 0.25$ on FEP), indicating complementary predictions. The anatomically-coordinated sensitivity profile is the cross-seed reproducible one.

The low-rank subspace is our modeling choice. The deformation carries classification and targets explicit ROI $\times$ frequency $\times$ lag coordinates, so its decision is attributable to anatomically labeled axes. The attribution claim rests on a cross-seed reproducible sensitivity axis (MDD, FEP) and highest-importance coordinates (cross-seed stable for all three tasks) aligned with the implicated circuitry, and we treat it as hypothesis-generating. Read this way, the representation returns a short list of ROI–band–lag coordinates that constrain where mechanistic hypotheses can be tested. This paper solves the estimation side of the closed-loop problem: $EEG_t \rightarrow p_{\mathrm{disease},t} + \Delta h_t$, leaving the stimulation-response model $\Delta h_{t+\Delta} = F(\Delta h_t, u_t)$ to future TMS-EEG work (transcranial magnetic stimulation with concurrent EEG). Our contribution is structured state estimation.

5.2 Limitations

The clinical cohorts are small and sLORETA constrains deep structures, so the anatomical claims are mesoscale by design: the 46 ROIs are shared source coordinates, adjacent-ROI differences fall short of millimeter-level localization, and basal-ganglia subnuclei are excluded. Lag coordinates are autocorrelation lags with a magnitude-only reading, and the communication readout is a model-implied predictive quantity, stopping short of synaptic conduction delays or a measured causal pathway. Coordinate-level localization relies on the internal stability tests (cross-fold/cross-seed stability, synthetic recovery, and lag perturbation; Appendix I.7) and on external medical-consistency checks. The fusion uses a fixed α=0.5, and it matches the best single branch within seed noise, with its advantage concentrated on the weaker branch. Attributions in Table 14 are directional decision-sensitivity findings, and the exact coordinate map is a directional result, with a validated signature left to larger cohorts.

6 Conclusion

We separate diagnosis from an anatomically-coordinated disease representation in clinical EEG: a simple scalp expert preserves diagnostic evidence, while a healthy-dynamics-conditioned low-rank deformation model learns an independently predictive source-space state in explicit ROI $\times$ frequency $\times$ lag coordinates. Under a strict subject-level protocol, decision-level fusion matches the best branch on MDD and FEP within seed noise and exceeds the source branch on PD. The source branch is the strongest single expert on FEP and direction-reproducible on MDD and FEP. Its ROI–frequency–lag attribution is concordant with known disease circuitry and, for MDD, transfers to a cohort recorded with a different montage. Because this reproducible, anatomically-indexed axis is explicit in anatomical coordinates, it is a structured state that future neuromodulation work could build on.

ACKNOWLEDGMENTS

We thank the MODMA project team and the OpenNeuro contributors of ds003944, ds004584, and the healthy pretraining corpora for making their data publicly available.

AI USE STATEMENT

We used generative AI tools in four roles: polishing writing; retrieving literature; supporting research execution; and drafting sections. *Research ideation and execution (required disclosure):* the research ideas, hypotheses, core method design, and key model code were developed by the human authors. AI tools assisted only with routine experimental and reproducibility code (data-processing, evaluation, plotting), feedback on the methodology and protocol, and interpretation of completed results. AI tools were unused for other required-disclosure tasks (synthetic data generation, proving mathematical claims). *Writing, retrieval, and drafting (recommended disclosure):* editing for readability, identifying literature, formatting references, suggesting structure, and proposing a title. All AI-assisted work was reviewed. *Elaboration: LLM suggestions for experimental design were checked against execution reports and result files before inclusion. LLM-assisted drafting and reference formatting were verified against the underlying data and the cited sources. Every AI-generated claim, number, and reference was author-verified.* We take responsibility for the final content of this work.

REPRODUCIBILITY STATEMENT

The protocol (non-overlapping multi-seed holdout, inner 3-fold checkpoint selection, one holdout evaluation per seed over $n{=}3$ seeds) is detailed in the appendix, with the hyperparameter table, the 46-ROI atlas, and the trajectory-stability rules. All experiments ran on one shared workstation with six custom-modified NVIDIA RTX 4090 GPUs (48 GB VRAM each). We use three stratified subject-split seeds, each with a nested subject-level cross-validation. Significance of the fusion gains is assessed by a subject-level bootstrap over the pooled holdout subjects, which anchors the variance estimate at the subject level. Source code is provided as supplementary material. A formal power analysis over larger multi-site cohorts is left to future work.

REFERENCES

Anonymous. LIGHT hypnotherapy. OpenNeuro, `https://doi.org/10.18112/openneuro.ds006437.v1.1.0`, 2025. Guided imagery hypnotherapy EEG dataset.

Luiz A. Baccalá and Koichi Sameshima. Partial directed coherence: a new concept in neural structure determination. *Biological Cybernetics*, 84(6):463–474, 2001. doi: 10.1007/PL00007990.

Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001. doi: 10.1023/A:1010933404324.

Hanshu Cai, Zhenqin Yuan, Yiwen Gao, Shuting Sun, Na Li, et al. A multi-modal open dataset for mental-disorder analysis. *Scientific Data*, 9(1):178, 2022. doi: 10.1038/s41597-022-01211-x.

James F. Cavanagh. Rest. OpenNeuro, `https://doi.org/10.18112/openneuro.ds003478.v1.0.0`, 2021. External resting-state MDD/HC cohort.

Di Chen, Xiang Wang, Valerie Voon, Yuchao Jiang, Chun-Yi Zac Lo, Linbo Wang, Chun Shen, Shitong Xiang, Shuqiao Yao, Jie Zhang, Tianye Jia, Wei Cheng, and Jianfeng Feng. Neurophysiological stratification of major depressive disorder by distinct trajectories. *Nature Mental Health*, 1:863–875, 2023. doi: 10.1038/s44220-023-00139-4.

Haoze Chen et al. Gumbel selection of sparse inputs with spectral and temporal fusion for EEG classification of parkinson's disease at the participant level. *Electronics*, 15(16):3563, 2026. doi: 10.3390/electronics15163563.

Hui Chen, Yanqin Lei, Rihui Li, Xinxin Xia, Nanyi Cui, Xianliang Chen, Jiali Liu, Huajia Tang, Jiawei Zhou, Ying Huang, Yusheng Tian, Xiaoping Wang, and Jiansong Zhou. Resting-state EEG dynamic functional connectivity distinguishes non-psychotic major depression, psychotic major

depression and schizophrenia. *Molecular Psychiatry*, 29(4):1088–1098, 2024. doi: 10.1038/s41380-023-02395-3.

Computational Neuroimaging and Neuroengineering Lab at the University of Oklahoma. Resting EEG. OpenNeuro, https://doi.org/10.18112/openneuro.ds006525.v1.0.0, 2025. Healthy resting-state high-density EEG dataset.

Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, 1995. doi: 10.1007/BF00994018.

Alena Damborská, Eliška Honzírková, Richard Barteček, Jana Hořínková, Sylvie Fedorová, Šimon Ondruš, Christoph M. Michel, and Maria Rubega. Altered directed functional connectivity of the right amygdala in depression: high-density EEG study. *Scientific Reports*, 10(1):4398, 2020. doi: 10.1038/s41598-020-61264-z.

Coralie de Hemptinne, Elena S. Ryapolova-Webb, Ellen L. Air, Paul A. Garcia, Kai J. Miller, Jeffrey G. Ojemann, Jill L. Ostrem, Nicholas B. Galifianakis, and Philip A. Starr. Exaggerated phase–amplitude coupling in the primary motor cortex in Parkinson disease. *Proceedings of the National Academy of Sciences*, 110(12):4780–4785, 2013. doi: 10.1073/pnas.1214546110.

Federico Del Pup, Louis Fabrice Tshimanga, Andrea Zanola, Luca Taffarello, Elisa Tentori, and Manfredo Atzori. Self-supervised representation learning for EEG-based detection of neurodegenerative diseases. *Applied Sciences*, 15(24):13275, 2025. doi: 10.3390/app152413275.

Lei Ding, Guofa Shou, Yoon-Hee Cha, John A. Sweeney, and Han Yuan. Brain-wide neural coactivations in resting human. *NeuroImage*, 260:119461, 2022. doi: 10.1016/j.neuroimage.2022.119461. URL https://doi.org/10.1016/j.neuroimage.2022.119461.

Sanaz Ghaffari, Arian Yavari, Sara Bonyadian, Arsalan Ghofrani, and Russell Butler. Visual EEG study (BrainVision → BIDS). OpenNeuro, https://doi.org/10.18112/openneuro.ds006547.v1.0.0, 2025. Healthy high-density visual stimulation EEG dataset.

Clive W. J. Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica*, 37(3):424–438, 1969. doi: 10.2307/1912791.

Fatemeh Hasanzadeh, Maryam Mohebbi, and Reza Rostami. Graph theory analysis of directed functional brain networks in major depressive disorder based on EEG signal. *Journal of Neural Engineering*, 17(2):026010, 2020. doi: 10.1088/1741-2552/ab7613.

Stefan Haufe, Frank Meinecke, Kai Görgen, Sven Dähne, John-Dylan Haynes, Benjamin Blankertz, and Felix Biessmann. On the interpretation of weight vectors of linear models in multivariate neuroimaging. *NeuroImage*, 87:96–110, 2014. doi: 10.1016/j.neuroimage.2013.10.067.

Aron T. Hill, Neil W. Bailey, Talitha C. Ford, and Jarrad A. G. Lum. Multivariate classification of first-episode schizophrenia spectrum psychosis using EEG microstate dynamics. *medRxiv*, 2026. doi: 10.64898/2026.02.18.26346582. URL https://www.medrxiv.org/content/10.64898/2026.02.18.26346582v1.

Bo Hjorth. EEG analysis based on time domain properties. *Electroencephalography and Clinical Neurophysiology*, 29(3):306–310, 1970. doi: 10.1016/0013-4694(70)90143-4.

Lu Huang, Zheng Hu, Zhengnan Zhang, and Yunyuan Gao. Alzheimer's disease detection using combined EEG source connectivity and microstate features. *Brain Sciences*, 16(8):856, 2026. doi: 10.3390/brainsci16080856.

Alberto Jaramillo-Jimenez, Diego A. Tovar-Rios, Johann Alexis Ospina, Yorguin-Jose Mantilla-Ramos, Daniel Loaiza-López, Verónica Henao Isaza, Luisa María Zapata Saldarriaga, Valeria Cadavid Castro, Jazmin Ximena Suarez-Revelo, Yamile Bocanegra, Francisco Lopera, David Antonio Pineda-Salazar, Carlos Andrés Tobón Quintero, John Fredy Ochoa-Gomez, Miguel Germán Borda, Dag Aarsland, Laura Bonanni, and Kolbjørn Brønnick. Spectral features of resting-state EEG in Parkinson's disease: A multicenter study using functional data analysis. *Clinical Neurophysiology*, 151:28–40, 2023. doi: 10.1016/j.clinph.2023.03.363.

Wei-Bang Jiang, Li-Ming Zhao, and Bao-Liang Lu. Large brain model for learning generic representations with tremendous EEG data in BCI. In *International Conference on Learning Representations (ICLR)*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/47393e8594c82ce8fd83adc672cf9872-Abstract-Conference.html.

Aya Kabbara, Gabriel Robert, Mahmoud Khalil, Marc Vérin, Pascal Benquet, and Mahmoud Hassan. An electroencephalography connectome predictive model of major depressive disorder severity. *Scientific Reports*, 12(1):6816, 2022. doi: 10.1038/s41598-022-10949-8.

Roselinde H. Kaiser, Jessica R. Andrews-Hanna, Tor D. Wager, and Diego A. Pizzagalli. Large-scale network dysfunction in major depressive disorder: A meta-analysis of resting-state functional connectivity. *JAMA Psychiatry*, 72(6):603–611, 2015. doi: 10.1001/jamapsychiatry.2015.0071.

M. J. Kamiński and K. J. Blinowska. A new method of the description of the information flow in the brain structures. *Biological Cybernetics*, 65(3):203–210, 1991. doi: 10.1007/BF00198091.

Andy Jeesu Kim, Mara Mather, Santiago Morales, and Joshua Senior. Young adult resting state and auditory oddball task EEG data. OpenNeuro, https://doi.org/10.18112/openneuro.ds006480.v1.0.1, 2025a. Healthy young-adult resting-state and auditory oddball EEG dataset.

Andy Jeesu Kim, Santiago Morales, Joshua Senior, and Mara Mather. HeartBEAM: Older adult resting state and auditory oddball task EEG data. OpenNeuro, https://doi.org/10.18112/openneuro.ds006466.v1.0.1, 2025b. Healthy older-adult resting-state and auditory oddball EEG dataset.

Isaac Kinley and Sue Becker. Cueing the future to reduce temporal discounting. OpenNeuro, https://doi.org/10.18112/openneuro.ds006446.v1.0.0, 2025. Healthy episodic future thinking and intertemporal decision EEG dataset.

Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 5338–5348, 2020. URL https://proceedings.mlr.press/v119/koh20a.html.

Vernon J. Lawhern, Amelia J. Solon, Nicholas R. Waytowich, Stephen M. Gordon, Chou P. Hung, and Brent J. Lance. EEGNet: A compact convolutional neural network for EEG-based brain–computer interfaces. *Journal of Neural Engineering*, 15(5):056013, 2018. doi: 10.1088/1741-2552/aace8c.

Haijun Lin et al. DEP-TFDualNet: a dual-domain attention framework with temporal-frequency fusion for depression recognition using three-channel frontal EEG. *Sensors*, 26(12):3861, 2026. doi: 10.3390/s26123861.

Vladimir Litvak, Anamika Jha, Alexandre Eusebio, Robert Oostenveld, Thomas Foltynie, Patricia Limousin, Ludvic Zrinzo, Marwan Hariz, Karl Friston, and Peter Brown. Resting oscillatory cortico-subthalamic connectivity in patients with Parkinson's disease. *Brain*, 134(Pt 2):359–374, 2011. doi: 10.1093/brain/awq332.

Shuang Liu, Sitong Chen, Zhenni Huang, Xiaoya Liu, Meijuan Li, Fangyue Su, Xinyu Hao, and Dong Ming. Hypofunction of directed brain network within alpha frequency band in depressive patients: a graph-theoretic analysis. *Cognitive Neurodynamics*, 16(5):1059–1071, 2022. doi: 10.1007/s11571-022-09782-6.

Muriel Lobier, Felix Siebenhühner, Satu Palva, and J. Matias Palva. Phase transfer entropy: a novel phase-based measure for directed connectivity in networks coupled by oscillatory interactions. *NeuroImage*, 85(Pt 2):853–872, 2014. doi: 10.1016/j.neuroimage.2013.08.056.

Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Raetsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 4114–4124, 2019. URL https://proceedings.mlr.press/v97/locatello19a.html.

Andre F. Marquand, Iead Rezek, Jan Buitelaar, and Christian F. Beckmann. Understanding heterogeneity in clinical cohorts using normative models: beyond case-control studies. *Biological Psychiatry*, 80(7):552–561, 2016. doi: 10.1016/j.biopsych.2015.12.023.

Andre F. Marquand, Seyed Mostafa Kia, Mariam Zabihi, Thomas Wolfers, Jan K. Buitelaar, and Christian F. Beckmann. Conceptualizing mental disorders as deviations from normative functioning. *Molecular Psychiatry*, 24:1415–1424, 2019. doi: 10.1038/s41380-019-0441-1.

María Melcón, Enrique Stern, Lydia Arana, and Almudena Capilla. EEG data during three near-threshold visual detection tasks: a no-cue task, a noninformative cue task (50% validity), and an informative cue task (100% validity). OpenNeuro, `https://doi.org/10.18112/openneuro.ds006171.v1.0.0`, 2025. Healthy near-threshold visual detection EEG dataset.

John Myers, Jiayang Xiao, Raissa K. Mathura, Ben Shofty, Victoria Gates, Joshua Adkinson, Anusha B. Allawala, Adrish Anand, Ron Gadot, Ricardo Najera, Hernan G. Rey, Sanjay J. Mathew, Kelly Bijanki, Garrett Banks, Andrew Watrous, Eleonora Bartoli, Sarah R. Heilbronner, Nicole Provenza, Wayne K. Goodman, Nader Pouratian, Benjamin Y. Hayden, and Sameer A. Sheth. Intracranial directed connectivity links subregions of the prefrontal cortex to major depression. *Nature Communications*, 16(1):6309, 2025. doi: 10.1038/s41467-025-61487-6.

Aisling OŃeill, Andrea Mechelli, and Sagnik Bhattacharyya. Dysconnectivity of large-scale functional networks in early psychosis: A meta-analysis. *Schizophrenia Bulletin*, 45(3):579–590, 2019. doi: 10.1093/schbul/sby094.

Jaehyun Park, Konyul Park, Daehun Kim, Junseo Park, and Jun Won Choi. Layer-wise modality decomposition for interpretable multimodal sensor fusion. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-4816. URL `https://proceedings.neurips.cc/paper_files/paper/2025/hash/d3e0aaa57c0d639f32a520cda39aec6d-Abstract-Conference.html`.

Roberto D. Pascual-Marqui. Standardized low-resolution brain electromagnetic tomography (sLORETA): technical details. *Methods and Findings in Experimental and Clinical Pharmacology*, 24(Suppl D):5–12, 2002.

Sadi Md. Redwan, Md Palash Uddin, Anwaar Ulhaq, Muhammad Imran Sharif, and Govind Krishnamoorthy. Power spectral density-based resting-state EEG classification of first-episode psychosis. *Scientific Reports*, 14(1):15154, 2024. doi: 10.1038/s41598-024-66110-0.

Dean Salisbury, Dylan Seebold, and Brian Coffman. EEG: First episode psychosis vs. control resting task 1. OpenNeuro, `https://doi.org/10.18112/openneuro.ds003944.v1.0.1`, 2022a. First-episode psychosis resting-state EEG cohort.

Dean Salisbury, Dylan Seebold, and Brian Coffman. EEG: First episode psychosis vs. control resting task 2. OpenNeuro, `https://doi.org/10.18112/openneuro.ds003947.v1.0.1`, 2022b. External first-episode psychosis resting-state cohort.

Robin Tibor Schirrmeister, Jost Tobias Springenberg, Lukas Dominique Josef Fiederer, Martin Glasstetter, Katharina Eggensperger, Michael Tangermann, Frank Hutter, Wolfram Burgard, and Tonio Ball. Deep learning with convolutional neural networks for EEG decoding and visualization. *Human Brain Mapping*, 38(11):5391–5420, 2017. doi: 10.1002/hbm.23730.

Lianne Schmaal, Dick J. Veltman, Theo G. M. van Erp, et al. Subcortical brain alterations in major depressive disorder: findings from the ENIGMA major depressive disorder working group. *Molecular Psychiatry*, 21(6):806–812, 2016. doi: 10.1038/mp.2015.69.

Yvette I. Sheline, Po W. Wang, Mokhtar H. Gado, John G. Csernansky, and Michael W. Vannier. Hippocampal atrophy in recurrent major depression. *Proceedings of the National Academy of Sciences*, 93(9):3908–3913, 1996. doi: 10.1073/pnas.93.9.3908.

Miseon Shim, Han-Jeong Hwang, Do-Won Kim, Seung-Hwan Lee, and Chang-Hwan Im. Machine-learning-based diagnosis of schizophrenia using combined sensor-level and source-level EEG features. *Schizophrenia Research*, 176(2–3):314–319, 2016. doi: 10.1016/j.schres.2016.05.007.

Arun Singh, Rachel Cole, Arturo Espinoza, Jim Cavanagh, and Nandakumar Narayanan. Rest eyes open. OpenNeuro, https://doi.org/10.18112/openneuro.ds004584.v1.0.0, 2024. Parkinson's disease resting-state EEG cohort.

Shuting Sun, Jianxiu Li, Huayu Chen, Tao Gong, Xiaowei Li, and Bin Hu. A study of resting-state EEG biomarkers for depression recognition. *arXiv preprint arXiv:2002.11039*, 2020. URL https://arxiv.org/abs/2002.11039.

Xiaoyu Tong, Hua Xie, Wei Wu, Corey J. Keller, Gregory A. Fonzo, Matthieu Chidharom, Nancy B. Carlisle, Amit Etkin, and Yu Zhang. Individual deviations from normative electroencephalographic connectivity predict antidepressant response. *Journal of Affective Disorders*, 351:220–230, 2024. doi: 10.1016/j.jad.2024.01.177.

Poul Videbech and Barbara Ravnkilde. Hippocampal volume and depression: a meta-analysis of MRI studies. *American Journal of Psychiatry*, 161(11):1957–1966, 2004. doi: 10.1176/appi.ajp.161.11.1957.

Jiquan Wang, Sha Zhao, Zhiling Luo, Yangxuan Zhou, Haiteng Jiang, Shijian Li, Tao Li, and Gang Pan. CBraMod: a criss-cross brain foundation model for EEG decoding. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/bbbd6d915cb90be21c1254a82d45cedd-Abstract-Conference.html.

Jiquan Wang, Sha Zhao, Yangxuan Zhou, Yiming Kang, Shijian Li, and Gang Pan. DeeperBrain: a neuro-grounded EEG foundation model towards universal BCI. *arXiv preprint arXiv:2601.06134*, 2026. URL https://arxiv.org/abs/2601.06134.

Peter D. Welch. The use of fast Fourier transform for the estimation of power spectra: a method based on time averaging over short, modified periodograms. *IEEE Transactions on Audio and Electroacoustics*, 15(2):70–73, 1967. doi: 10.1109/TAU.1967.1161901.

Alexis E. Whitton, Stephanie Deccy, Manon L. Ironside, Poornima Kumar, Miranda Beltzer, and Diego A. Pizzagalli. Electroencephalography source functional connectivity reveals abnormal high-frequency communication among large-scale functional networks in depression. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, 3(1):50–58, 2018. doi: 10.1016/j.bpsc.2017.07.001.

Ying Choon Wu, Enrique Carrillosulub, Leon Lange, Nicole Wells, Paula Jackson, Thomas Thudiyanplackal, and Paul J. Mills. Alleviating job-related burnout in medical professionals through guided imagery hypnotherapy. *Journal of Evidence-Based Integrative Medicine*, 31:1–11, 2026. doi: 10.1177/2515690X261418396. URL https://doi.org/10.1177/2515690X261418396.

Jingru Yang, Bowen Li, Wanqing Dong, Xiaorong Gao, and Yanfei Lin. Time-varying EEG networks of major depressive disorder during facial emotion tasks. *Cognitive Neurodynamics*, 18(5):2605–2619, 2024. doi: 10.1007/s11571-024-10111-2.

Alexandra L. Young, Razvan V. Marinescu, Neil P. Oxtoby, Martina Bocchetta, Keir Yong, Nicholas C. Firth, David M. Cash, David L. Thomas, Katrina M. Dick, Jorge Cardoso, et al. Uncovering the heterogeneity and temporal complexity of neurodegenerative diseases with subtype and stage inference. *Nature Communications*, 9(1):4273, 2018. doi: 10.1038/s41467-018-05892-0.

A Evaluation protocol details

A.1 Leak-free multi-seed holdout

For each seed $s \in \{42, 99, 2026\}$: a stratified, disjoint 20% subject holdout is locked as the test set (the three seeds' holdouts are disjoint). The remaining pool is split into 3 subject-level stratified fold pairs (train 2/3, val 1/3). Each fold trains with early stopping and stores its fold-best checkpoint by a subject-level metric. The inner-best checkpoint is selected from the three fold-best checkpoints by

Algorithm 1 Neural-dynamics-conditioned iterative disease deformation

Require: Source representation, healthy model f_{θ_H}, shared directions V_D, amplitude network s, steps K, top-m
Ensure: Deformation state Δh^K, diagnosis $\hat{y}^K$
$h^0 \leftarrow f_{\theta_H}(x)$ ▷ self-dynamics anchor
$\Delta h^0 \leftarrow 0$
for $k = 0, \dots, K-1$ **do**
  $U_k(x) \leftarrow V_D^\top s_k(x)$ ▷ low-rank disease direction, per-subject signed amplitude
  $\eta_k \leftarrow \eta_{\min} + (\eta_{\max} - \eta_{\min})\ \mathrm{sigmoid}(\tilde{\eta}_k);\ \tau_k \leftarrow \tau_{\min} + (\tau_{\max} - \tau_{\min})\ \mathrm{sigmoid}(\tilde{\tau}_k)$
  $\tilde{\Delta} h^{k+1} \leftarrow \Delta h^k - \eta_k U_k(x)$
  $\Delta h^{k+1} \leftarrow \mathrm{TopK}_m\big[\,\mathrm{SoftThreshold}(\tilde{\Delta} h^{k+1}, \tau_k)\big]$
  $\hat{y}^k \leftarrow \mathrm{Head}_D(\Delta h^{k+1})$ ▷ delta_h readout
end for
return $\Delta h^K,\ \hat{y}^K$

inner-validation subject AUC (ties broken by the selection metric), and that one model is evaluated once on the seed's holdout. The reported value is mean±std across the 3 seeds (n=3 independent holdout evaluations); inner-fold scores serve only as selection references. Holdout subjects are excluded from every fold's training and validation, and subject-level clip grouping guarantees no clip of a test subject appears in training.

We use non-overlapping subject partitions by design and report the variance. Statistical significance is assessed via a subject-level bootstrap, independent of the number of seeds.

A.2 Model selection metric

Two selection stages use different metrics. Early stopping and the fold-best checkpoint use a subject-level metric, by default the subject-level negative log-likelihood (clip probabilities aggregated per subject). This stabilizes MDD across seeds but is sometimes suboptimal (FEP can prefer clip-level selection), so it is treated as a per-disease setting. The inner-best checkpoint across the three folds, and the solver depth and rank in the K/r scan, are then selected by inner-validation subject AUC.

A.3 Trajectory reporting rules

A layer k^* is eligible for discussion only if the trajectory type is stable-accumulating or final-layer-selective. Oscillatory/reversing trajectories are marked "unstable_oscillatory" and are withheld from biological interpretation. For every layer we report direction similarity between adjacent Δh steps, per-layer $|\Delta h|$, saturation rate, and whether candidate coordinates flip sign across $k = 1..K$.

A.4 Deformation solver as an algorithm

Algorithm 1 summarizes the iterative deformation of Section 3.4: the per-step update depends only on the source representation, the anchor h^0, and the current deformation, with the disease label entering only the outer objective Eq. 7.

B Preprocessing and leakage control

B.1 Signal preprocessing

All signals are band-pass filtered (0.5–70 Hz) with a 50 Hz notch, re-referenced to the common average, and source-reconstructed with sLORETA on a shared template head model into the 46-ROI atlas. For the seven healthy pretraining corpora (Section 4.1; 269 subjects), to bound the compute budget and balance subjects across cohorts, each subject contributes up to 20 randomly sampled 10 s segments (50% overlap), with subject-level grouping enforced. Preprocessing parameters are fixed per dataset family before any split is drawn, so no test-subject statistics enter preprocessing.

B.2 LEAKAGE CONTROL IN CROSS-ROI ANALYSES

Whenever cross-ROI predictive increments or direction asymmetries are reported, we apply symmetric orthogonalization of the ROI time series to remove zero-lag common-driver components before recomputing the quantity, and we report both raw and leakage-corrected values (see the supplementary source-side analyses). This control is decisive: raw cross-ROI increments disappear after correction, so no claim of directed communication is made from increments alone.

C BASELINE REPRODUCTION DETAILS

All learning-based baselines share the proposed model's subject-level protocol: the same non-overlapping multi-seed holdout, subject-level metric aggregation (clip probabilities averaged per subject), and the restriction that holdout subjects never enter training, early stopping, or model selection. EEGNet is trained from random initialization on the four-band filtered series at the native montage following its published architecture, and its source variant replaces the input with the 46-ROI source series; CBraMod uses the authors' official checkpoint and fine-tuning pipeline; DeeperBrain follows its publication's protocol, as its training code was unavailable, and its near-chance subject-level results reflect the difficulty of subject-wise generalization on these small cohorts. On MDD and FEP several learning-based baselines collapse to a single predicted class, giving a balanced accuracy of exactly 0.5 with zero variance; their AUC stays above 0.5, so this is a threshold degeneracy at these cohort sizes, distinct from a numerical failure. The microstate baseline follows the reported FEP microstate construction, and the PD spectral random forest follows (Chen et al., 2026). Reproduction choices (input scaling, segmentation length, montage alignment, checkpoint provenance) are fixed per baseline and listed with the shared settings in the result records.

D SCALP VERSUS SOURCE SPECTRAL DISCRIMINATION

Table 4 applies two probes to the scalp and to the 46-ROI source representation: a per-disease PSD probe (the best of logistic regression, random forest, and SVM) and EEGNet (Lawhern et al., 2018). The penalty is source minus scalp within the same probe (negative $=$ weaker source). Comparing the probes tests whether a higher-capacity source classifier closes the gap.

Table 4: Representation dilemma under the strict protocol (AUC). Each method is applied to both representations: the PSD probe uses, for each disease, the best of logistic regression, random forest, and SVM on the PSD features, applied identically to the scalp and source-ROI PSD, and EEGNet (Lawhern et al., 2018) is trained from random initialization on the scalp channels and on the 46-ROI source time series under the identical holdout. The penalty is source minus scalp within the same method; a negative penalty means the source representation is weaker. Values are subject-level mean AUC from the same runs as Table 2.

Disease	Method	Scalp	Source	Penalty
MDD	PSD	0.667	0.589	-0.078
MDD	EEGNet	0.522	0.533	$+0.011$
FEP	PSD	0.700	0.533	-0.167
FEP	EEGNet	0.639	0.628	-0.011
PD	PSD	0.803	0.640	-0.163
PD	EEGNet	0.743	0.587	-0.156

E ROI ATLAS AND HYPERPARAMETERS

E.1 ROI ATLAS (46)

The atlas is fixed before disease-model evaluation. Each left/right entry below is one separate coordinate, giving $23 \times 2 = 46$ ROIs.

Table 5: Complete fixed 46-ROI atlas. L and R are distinct source coordinates. The names are the analysis labels used throughout the code and result records. Design groups record the inclusion rationale fixed before disease-model evaluation.

Pair	ROI labels (two coordinates)	Design group	Anatomical name
1	DLPFC_L, DLPFC_R	MDD core	Dorsolateral prefrontal cortex
2	dmPFC_L, dmPFC_R	MDD core	Dorsomedial prefrontal cortex
3	lateral_OFC_L, lateral_OFC_R	MDD core	Lateral orbitofrontal cortex
4	medial_OFC_L, medial_OFC_R	MDD core	Medial orbitofrontal cortex
5	dACC_L, dACC_R	MDD core	Dorsal anterior cingulate cortex
6	sgACC_L, sgACC_R	MDD core	Subgenual anterior cingulate cortex
7	PCC_L, PCC_R	MDD core	Posterior cingulate cortex
8	insula_L, insula_R	MDD core	Insular cortex
9	amygdala_L, amygdala_R	MDD core	Amygdala
10	hippocampus_L, hippocampus_R	MDD core	Hippocampus
11	thalamus_L, thalamus_R	MDD core	Thalamus
12	IFG_L, IFG_R	FEP language/attention	Inferior frontal gyrus
13	STG_auditory_L, STG_auditory_R	FEP language/attention	Auditory superior temporal gyrus
14	MTG_L, MTG_R	FEP language/attention	Middle temporal gyrus
15	precuneus_L, precuneus_R	FEP language/attention	Precuneus
16	SPL_L, SPL_R	FEP language/attention	Superior parietal lobule
17	IPL_angular_L, IPL_angular_R	FEP language/attention	Angular part of inferior parietal lobule
18	M1_L, M1_R	PD motor	Primary motor cortex
19	premotor_L, premotor_R	PD motor	Premotor cortex
20	SMA_L, SMA_R	PD motor	Supplementary motor area
21	S1_L, S1_R	PD motor	Primary somatosensory cortex
22	parahippocampal_L, parahippocampal_R	Robustness extension	Parahippocampal cortex
23	visual_association_L, visual_association_R	Robustness extension	Visual association cortex

The 22 MDD-core coordinates were expanded before final evaluation by 12 FEP language/attention, 8 PD motor, and 4 robustness coordinates. Basal-ganglia subnuclei are excluded as primary ROIs: template-head sLORETA with routine scalp EEG falls short of supporting independent subnuclear claims.

E.2 HYPERPARAMETERS

Table 6: Default hyperparameters.

Parameter	Value
ROI count	46
Bands	$\delta, \theta, \alpha, \beta, \gamma$
Lag indices	4 (real per-τ ACF, $\tau \in \{1, 2, 4, 8\}$)
K (iterations)	4 (MDD), 3 (FEP), 2 (PD)
Low-rank dimension r	1 (MDD/FEP), 2 (PD)
delta_topk m	460 (top-half of 920)
Amplitude bound $s_{\max}$	10
λ (ℓ_1 on Δh^K)	0.01
Direction/amplitude MLP width	64
OP-CNN	d_{model}=64; stem/block kernels $7/3$; dilations $(1, 2, 4, 8)$
Head (with FL-DPD)	MLP 42320→256→2
Head (source)	linear 920→2
Batch size	4
Optimizer	AdamW, cosine-annealing LR schedule
Early stopping patience	5

The full diagnostic path is compact: the source branch has 168k trainable parameters (OP-CNN 104k; deformation solver and direction net 60k; linear readout 1.8k), the sensor expert adds its logistic-regression weights, and the total DMD-EEG model has $\approx$ 169k parameters (MDD 169,628; FEP 168,623; PD 169,592). The g_{norm} development readout is excluded from the reported results.

Table 7: Solver depth and rank scan on the real-lag scale-axis (per-τ ACF) under the strict three-seed protocol (mean$\pm$std over seeds), including the one-shot ($K{=}1$) and no-deformation ($K{=}0$) controls. Candidates are the $K{\geq}2$ rows; the bold row is the inner-validation argmax over its disease's candidates. The control rows ($K{=}1$ and $K{=}0$) fall outside the candidate set. *The no-deformation ($K{=}0$) row bypasses the solver entirely and routes the anchor state h^0 directly to the linear head, so it isolates the contribution of the iterative deformation beyond a zero-input control. The holdout columns are a post-selection check.

Disease	Solver	Inner-val AUC	Holdout AUC	Holdout BAcc
MDD	$K{=}4, r{=}1$	$\mathbf{0.801 \pm 0.091}$	0.622 ± 0.134	0.622 ± 0.110
MDD	$K{=}3, r{=}1$	0.794 ± 0.060	0.544 ± 0.096	0.550 ± 0.049
MDD	$K{=}4, r{=}2$	0.785 ± 0.126	0.622 ± 0.063	0.567 ± 0.121
MDD	$K{=}1, r{=}1$	0.785 ± 0.252	0.567 ± 0.119	0.506 ± 0.021
MDD	$K{=}0$ (no deformation)*	0.780 ± 0.085	0.600 ± 0.054	0.594 ± 0.031
FEP	$K{=}4, r{=}1$	0.719 ± 0.086	0.628 ± 0.031	0.567 ± 0.027
FEP	$K{=}3, r{=}1$	$\mathbf{0.826 \pm 0.110}$	0.794 ± 0.139	0.706 ± 0.131
FEP	$K{=}1, r{=}1$	0.796 ± 0.071	0.667 ± 0.142	0.611 ± 0.172
FEP	$K{=}0$ (no deformation)*	0.698 ± 0.108	0.678 ± 0.110	0.578 ± 0.075
PD	$K{=}4, r{=}1$	0.673 ± 0.051	0.552 ± 0.090	0.550 ± 0.061
PD	$K{=}2, r{=}2$	$\mathbf{0.680 \pm 0.051}$	0.638 ± 0.020	0.608 ± 0.077
PD	$K{=}1, r{=}1$	0.664 ± 0.050	0.500 ± 0.108	0.508 ± 0.031
PD	$K{=}0$ (no deformation)*	0.651 ± 0.034	0.623 ± 0.049	0.550 ± 0.041

Table 8: Healthy-dynamics pretraining ablation under the strict three-seed protocol (mean$\pm$std over seeds). "Full" is the reported configuration with the disease's selected solver depth and rank (Table 7); "scratch" trains the same configuration from an empty pretrained checkpoint, so the healthy-dynamics prior is removed. The $K{=}1$ and $K{=}0$ deformation controls are in Table 7.

Disease	Configuration	Inner-val AUC	Holdout AUC	Holdout BAcc
MDD	full (healthy pretrained)	0.801 ± 0.091	0.622 ± 0.134	0.622 ± 0.110
MDD	scratch (no pretraining)	0.756 ± 0.165	0.600 ± 0.027	0.578 ± 0.100
FEP	full (healthy pretrained)	0.826 ± 0.110	0.794 ± 0.139	0.706 ± 0.131
FEP	scratch (no pretraining)	0.758 ± 0.052	0.778 ± 0.129	0.667 ± 0.095
PD	full (healthy pretrained)	0.680 ± 0.051	0.638 ± 0.020	0.608 ± 0.077
PD	scratch (no pretraining)	0.634 ± 0.075	0.415 ± 0.076	0.508 ± 0.072

F Ablations

With the support fixed at $m{=}460$, the solver depth and rank are selected per disease by inner validation (Table 7): MDD $K{=}4, r{=}1$, FEP $K{=}3, r{=}1$, PD $K{=}2, r{=}2$. The cohorts differ in subject count, montage, and sampling rate, so we set the depth and rank per dataset over a shared ROI–band–lag coordinate system. The scan is deliberately small—the shared low-rank subspace is a capacity prior for small cohorts ($r \leq 2$) and the deformation is a short refinement ($K \leq 4$)—and both controls stay below the selected depth on every task, so the iterative deformation carries the discriminative signal beyond a one-shot low-rank projection or the undeformed anchor state. The scanned candidates differ by at most about one seed standard deviation on the selection metric, and the inner-validation–holdout gap (e.g., MDD $0.801 \rightarrow 0.622$) is the optimism of selecting among a few configurations on a small inner pool; the reported value is the single holdout evaluation per seed.

Clearing the pretrained checkpoint degrades the source branch on all three tasks, most sharply on PD: holdout AUC falls from 0.638 to 0.415 (PD), 0.794 to 0.778 (FEP), and 0.622 to 0.600 (MDD), with inner-validation AUC dropping on every task, so the healthy-dynamics prior stabilizes the learned deformation, doing more than initialization. The deformation's own contribution is the no-deformation ($K{=}0$) control in Table 7.

F.1 FUSION VARIANTS

We also evaluated two adaptive alternatives to the fixed average: a global α scanned on the inner validation fold (F1), and logit-space fusion $z_{\text{final}} = \alpha z_{\text{sensor}} + (1-\alpha) z_{\text{source}}$ with α chosen on the inner validation fold (F2). Neither reliably improved over the fixed average, and the selected α often collapsed to 0 because the inner pool is too small to estimate a stable weight. We therefore keep the fixed α=0.5 average, which also preserves the structural independence of the source state.

G PAIRED BOOTSTRAP OF THE FUSION GAIN

Table 9 reports the paired AUC difference between the fixed fusion and each single branch: fold-level values (descriptive) and subject-level values from a paired bootstrap over the pooled holdout subjects (20000 resamples, 95% percentile CI, same resampling index for both). The PD fusion-minus-source gain has a 95% CI that excludes zero.

Table 9: Paired ΔAUC of the α=0.5 fusion against each single branch. Fold-level rows report mean±std over the three holdout folds (folds with positive ΔAUC in parentheses). Subject-level rows report a paired bootstrap over the pooled holdout subjects (n=33/48/90; 20000 resamples; 95% percentile CI). Bold marks the comparison whose 95% CI excludes zero.

Disease	Level	ΔAUC Fusion−Source	ΔAUC Fusion−Sensor
MDD	fold (n=3)	$+0.078 \pm 0.096$ (2/3)	$+0.078 \pm 0.107$ (2/3)
MDD	subject (n=33)	+0.067 [−0.062, +0.198]	+0.052 [−0.070, +0.173]
FEP	fold (n=3)	-0.006 ± 0.139 (1/3)	$+0.089 \pm 0.111$ (2/3)
FEP	subject (n=48)	+0.030 [−0.104, +0.172]	+0.065 [−0.030, +0.164]
PD	fold (n=3)	$+0.142 \pm 0.026$ (3/3)	$+0.008 \pm 0.020$ (2/3)
PD	subject (n=90)	**+0.176 [+0.048, +0.308]**	−0.006 [−0.049, +0.035]

H FIXED FIVE-FOLD REFERENCE

For ablation control we report fixed subject-level five-fold results in Table 10. These values are selection-biased and the multi-seed holdout protocol replaces them as the trustworthy generalization estimate. Within this protocol, the best source-side configuration is the per-subject logit average of the two independently trained readouts, the direct-deformation and communication-pathway heads (MDD 0.763/0.712; FEP 0.683/0.693).

Table 10: Fixed subject-level five-fold reference, mean±std. Not the primary estimate. Ensemble rows (†) are per-subject logit averages of the two readouts. The strict-holdout results in the main text use the direct deformation readout. This 5-fold table additionally includes the communication-pathway readout as a development comparison. The no-deformation (K=0) control is reported in Table 7.

Disease	Readout	Acc	BAcc	AUC	F1
MDD	with FL-DPD	0.696 ± 0.111	0.705 ± 0.092	0.686 ± 0.158	0.676 ± 0.135
MDD	Ours (source)	0.642 ± 0.105	0.657 ± 0.109	0.698 ± 0.092	0.623 ± 0.116
MDD	Ensemble†	—	0.763†	0.712†	—
FEP	with FL-DPD	0.719 ± 0.073	0.660 ± 0.097	0.669 ± 0.099	0.674 ± 0.111
FEP	Ours (source)	0.682 ± 0.084	0.634 ± 0.094	0.683 ± 0.088	0.649 ± 0.111
FEP	Ensemble†	—	0.683†	0.693†	—
PD	with FL-DPD	0.671 ± 0.045	0.620 ± 0.083	0.598 ± 0.160	0.631 ± 0.059
PD	Ours (source)	0.624 ± 0.035	0.581 ± 0.043	0.613 ± 0.084	0.618 ± 0.034

I SUPPLEMENTARY SOURCE-SIDE ANALYSES

I.1 FL-DPD COMMUNICATION-PATHWAY ABLATION

To test whether the frequency–lag directed pathway contributes to source-side classification, we zero the pathway-fusion weights at each 5-fold best checkpoint and rerun inference on the same validation subjects. This drops subject-level accuracy from 0.611 ± 0.019 to 0.491 ± 0.046 and AUC from 0.666 ± 0.016 to 0.500, so the pathway carries most of the discriminative signal within this communication-pathway readout. The reported strict-holdout model reads the deformation Δh^K directly (Section 3.4), so this is a development comparison, separate from the reported decision path; it supports a predictive role for the pathway while leaving direction and lag semantics untested.

I.2 COMPONENT-LEVEL DIAGNOSTICS OF THE SOURCE MODEL

We examined which internal FL-DPD components train and carry information. In the earliest healthy-pretrained variants, the sparse gate (Hard-Concrete log-α) stayed at its initialization (mean open rate 0.043 over all $1{,}936$ gates), the cross-frequency term stayed near zero, and the scoring matrix S was nearly uniform across bands, with no single edge significant under subject-level permutation (0/462). A label-shuffle permutation test (5 permutations) confirmed the classification signal is real: observed AUC 0.69–0.74 versus null 0.458 ± 0.055. The source-side signal is therefore diffuse across edges and strongest in spectral power.

I.3 LEAKAGE CORRECTION FOR CROSS-ROI PREDICTIVE INCREMENTS

A negative control asks whether cross-ROI delayed prediction reflects directed communication or zero-lag common-driver leakage. Raw cross-ROI incremental predictive ability (ΔR^2, self AR(4) vs. self+cross AR(4)) was small ($+0.0116$ to $+0.0144$ across diseases) and vanished after symmetric orthogonalization (-0.0003), with the top-1 leakage component explaining 46–52% of variance and the top-3 explaining about 100% of the increment. Time reversal left this unchanged, and single-lag Granger increments and MVAR+PDC asymmetries also vanished. We therefore frame source-side communication as *model-implied predictive communication*, distinct from biological directionality. The FL-DPD ablation establishes a predictive contribution, leaving causal pathway identity open.

I.4 CROSS-TASK STATE DEPENDENCE OF MDD DEVIATIONS

The cross-task experiment from resting to the dot-probe event-related-potential (ERP) task (same 53 MODMA subjects, resting-trained model) found no false-discovery-rate (FDR)-significant frequency, lag, or direction edges in either state (resting 6/462 nominal frequency edges, ERP 7/462; best ERP lag p=0.409). The resting network centered on DMN nodes (sgACC–PCC–dmPFC–amygdala), the ERP state on a cognitive-control pattern (DLPFC–amygdala–hippocampus–thalamus). The pathway-by-state interaction failed to reach significance ($t = -0.985$, p=0.329, $d = -0.269$, n=53), and masking predefined DMN or fronto-limbic edge sets changed AUC by less than 0.01. These results support a preliminary state dependence, falling short of a verified dissociation, and constrain the single-state coordinate findings.

I.5 EXTERNAL CROSS-COHORT TRANSFER

The source branch is defined in the shared 46-ROI space and therefore uses the same source coordinate system across montages. The sensor expert and the fusion are tied to the recording montage and cannot be applied when the target cohort uses a different cap, so only the source branch is transferred. Table 11 reports the three-seed transfer for MDD (MODMA $\rightarrow$ ds003478 (Cavanagh, 2021)) and FEP (ds003944 $\rightarrow$ ds003947 (Salisbury et al., 2022b)).

For MDD the ranking transfers (AUC 0.724 ± 0.042; group-mean gap $+0.093 \pm 0.029$, positive in all three seeds), though the fixed 0.5 threshold mislabels many subjects (BAcc 0.548 ± 0.073), a calibration shift that leaves the signal intact. For FEP the group means are ordered correctly (gap $+0.045 \pm 0.024$) but the transfer AUC is near chance (0.542 ± 0.032), so the ds003944 discrimination is cohort-specific and we make no cross-cohort generalization claim for FEP. Both cohorts are small

Table 11: Cross-dataset transfer of the source branch (mean±std over the three seeds, each checkpoint evaluated once on the target cohort). The sensor expert and the fusion are tied to the recording montage and are excluded from transfer.

Transfer	Target cohort (n)	AUC	BAcc
MDD $\rightarrow$ ds003478	11 MDD / 75 HC	0.724 ± 0.042	0.548 ± 0.073
FEP $\rightarrow$ ds003947	31 psychosis / 30 control	0.542 ± 0.032	0.543 ± 0.020

and the folded metrics are class-balance sensitive, so only the ranking AUC is used as transfer evidence.

I.6 Additional ablations

Temporal position embeddings and attention-versus-mean pooling showed no reliable contribution: training only the position embedding and head changed MDD accuracy by $+0.056 \pm 0.046$ and AUC by -0.007 ± 0.013 (3/5 folds), and attention pooling was indistinguishable from mean pooling (AUC $+0.013 \pm 0.027$). In the FEP fine-tuning grid, the observed degradation traced to the focal-loss exponent (γ=2 $\rightarrow$ 0), while the communication loss weight stayed unchanged (fold-1 AUC $0.833 \rightarrow 0.500$). Discriminative power is thus concentrated in the spectral and deformation paths, with patch order and pooling contributing little.

I.7 Lag-axis validity: synthetic recovery and lag perturbation

The lag axis is the real per-τ autocorrelation of the op-CNN latent state ($\tau \in \{1, 2, 4, 8\}$), so its validity is established at the method level, independent of matching a lag to a literature value. Two experiments test the axis in opposite directions: a planted-lag recovery (positive control, whether the representation can encode a known lag) and a source-level lag perturbation (necessity, whether the decision uses it). The perturbations act on the solver's input state h_{health} of the chosen-fold checkpoint on the fold-validation subjects, and we report the clip-level AUC change relative to the identity perturbation.

Synthetic recovery. A narrowband oscillation at frequency $f_{\mathrm{hp}}/\ell^\star$ is planted in one target source channel so that its autocorrelation peaks at $\ell^\star \in \{1, 2, 4, 8\}$, and a logistic probe on the four-lag ACF predicts $\ell^\star$ with a half/half train/test split. Both the op-CNN latent ACF and the raw ACF recover the planted lag at accuracy 1.00 (chance 0.25; Table 12), confirming that the lag coordinates encode a known lag.

Table 12: Synthetic lag recovery (positive control): a logistic probe on the four-lag ACF recovers the planted lag $\ell^\star \in \{1, 2, 4, 8\}$ from either the op-CNN latent or the raw ACF (train/test split in half).

Probe input	Accuracy	Chance
op-CNN latent ACF	1.00	0.25
Raw ACF	1.00	0.25

Lag perturbation. Table 13 reports the three-seed mean ΔAUC at the solver input. Replacing the per-τ ACF by a time mean (the abstract-lag condition) gives the largest drop (MDD -0.160, FEP -0.138, PD -0.086), the same ordering as the real-lag versus abstract-lag comparison; permuting or rolling the τ anchoring gives smaller drops (-0.03 to -0.06). In the incremental test, lag adds beyond band on MDD (-0.034) and PD (-0.047), and fails to add on FEP ($+0.006$). The band and ROI perturbations are positive controls for the other axes.

The ACF is even in τ, so time reversal and time roll leave the four-lag features invariant ($\cos = 1.00$, $\Delta\mathrm{AUC} \approx 0$). The lag axis therefore encodes the magnitude $|\tau|$, leaving direction unsigned. A directed-lag claim would require an asymmetric lag feature.

Table 13: Source-level lag perturbation at the solver input (three-seed mean ΔAUC relative to the identity perturbation; chosen-fold checkpoint on fold-validation subjects). The incremental rows add lag on top of band or ROI, with the increment over the single-axis perturbation in parentheses.

Perturbation	MDD	FEP	PD
τ anchoring permuted	-0.043	-0.049	-0.052
τ anchoring rolled	-0.063	-0.034	-0.062
per-τ ACF $\rightarrow$ time mean	-0.160	-0.138	-0.086
full time shuffle	-0.074	-0.101	-0.163
band permuted (control)	-0.073	-0.132	-0.024
ROI permuted (control)	-0.124	-0.177	-0.079
lag + band	-0.106 (-0.034)	-0.126 ($+0.006$)	-0.072 (-0.047)
lag + ROI	-0.157 (-0.034)	-0.165 ($+0.011$)	-0.078 ($+0.001$)

Table 14: **Attribution readout of the healthy-anchored state h_{health}.** ROI importance is the unsigned amplitude saliency $\text{mean}_{\text{subj}}|\partial \,\text{logit}/\partial h_{\text{health}}|$ (consensus over 3 splits $\times$ 3 folds), cross-seed stable (axis cosine 0.83–0.87) and concordant with the occlusion test below (ROI Spearman $+0.31$ MDD, $+0.59$ PD). The spectro-temporal signature is the ROI-marginal signed saliency profile, carrying band and lag direction and selectivity. Each coordinate sits on a real axis: ROI and band inherited from the input, lag the real per-τ autocorrelation ($\tau \in \{1, 2, 4, 8\}$). A band/lag $(+)$ sign means the patient-versus-healthy logit rises as that coordinate increases, marking a direction of decision sensitivity, distinct from a change in brain activity. Directional saliency findings.

Disease	High-importance ROIs (amplitude)	Spectro-temporal signature (signed)
MDD	MTG_R, DLPFC_R, parahippocampal_L, amygdala_R, dACC_R, dmPFC_L, sgACC_L, PCC_L	**Band:** $\delta\ -1.40 \cdot \theta\ +0.73 \cdot \alpha\ +0.77 \cdot \beta\ -0.72 \cdot \gamma\ +0.33$ **Lag:** $\tau_1\ +0.57 \cdot \tau_2\ -0.59 \cdot \tau_4\ +0.02 \cdot \tau_8\ -0.29$
FEP	IFG_R, visual_association_R, precuneus_R, SPL_L, insula_R, IPL_angular_R, dmPFC_R, sgACC_L	**Band:** $\delta\ +0.26 \cdot \theta\ -0.24 \cdot \alpha\ -0.45 \cdot \beta\ +0.31 \cdot \gamma\ -0.20$ **Lag:** $\tau_1\ -0.84 \cdot \tau_2\ -0.12 \cdot \tau_4\ -0.03 \cdot \tau_8\ +0.66$
PD	premotor_L, IPL_angular_R, IPL_angular_L, S1_L, M1_L, lateral_OFC_R, SMA_R, precuneus_L	**Band:** $\delta\ -0.13 \cdot \theta\ +0.01 \cdot \alpha\ +0.75 \cdot \beta\ -1.58 \cdot \gamma\ +0.01$ **Lag:** $\tau_1\ -0.13 \cdot \tau_2\ -0.14 \cdot \tau_4\ -0.11 \cdot \tau_8\ -0.55$

I.8 Attribution coordinates

The consolidated highest-importance coordinates discussed in the main text are listed in Table 14.

I.9 ROI occlusion on the decision

Table 15 occludes the top-ranked ROIs and joint ROI sets: the set's full band $\times$ lag block(s) in h_{health} are zeroed and the holdout re-scored. Single ROIs are ranked by occlusion magnitude over all 46 ROIs, independent of saliency; joint sets (†) are occlusion-search results, reported as exploratory.

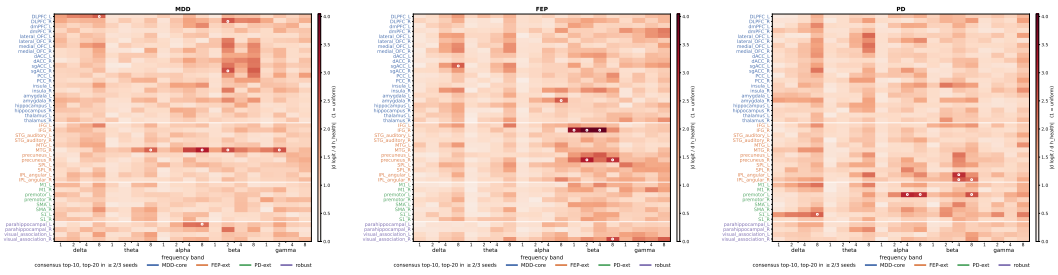
J Supplementary figures

These figures are descriptive (only the AUROC panels give unbiased performance estimates) and show *where and in which direction* the source state differs.

Table 15: ROI occlusion on the healthy-anchored state. AUC drop (mean±std over the three seed checkpoints, chosen fold) when the occluded set's band × lag block is zeroed and the holdout is re-scored. Single ROIs are ranked by occlusion magnitude over all 46 ROIs; joint sets (†) are occlusion-search results and exploratory. The permutation p over random same-size block sets (internal null, n=300) is nominal and uncorrected for the search over 46 ROIs, so the table is a descriptive ranking only.

Disease	Occluded set	ΔAUC ± std (p)
FEP	SPL_L	$+0.056 \pm 0.042$ (0.003)
	SPL_R	$+0.050 \pm 0.049$ (0.004)
	dmPFC_L	$+0.050 \pm 0.014$ (0.005)
	insula_R	$+0.044 \pm 0.028$ (0.035)
	M1_R	$+0.039 \pm 0.016$ (0.017)
	thalamus_R	$+0.033 \pm 0.036$ (0.018)
	visuo_motor6†	$+0.106$ (0.040)
PD	premotor_L	$+0.028 \pm 0.016$ (0.004)
	M1_L	$+0.023 \pm 0.024$ (0.010)
	dACC_L	$+0.017 \pm 0.010$ (0.028)
	motor_insula3†	$+0.030$ (0.060)
MDD	dACC_R	$+0.044 \pm 0.016$ (0.045)
	motor_dacc6†	$+0.067$ (0.020)
	limbic6†	$+0.056$ (0.060)

† motor_dacc6 = {IFG_L, M1_L, SMA_L, amygdala_R, dACC_R, premotor_L}; limbic6 = {DLPFC_L, amygdala_R, dACC_L, hippocampus_L, parahippocampal_L, thalamus_L}; visuo_motor6 = {SPL_L, SPL_R, STG_auditory_R, premotor_L, premotor_R, sgACC_R}; motor_insula3 = {insula_L, precuneus_R, premotor_L}.



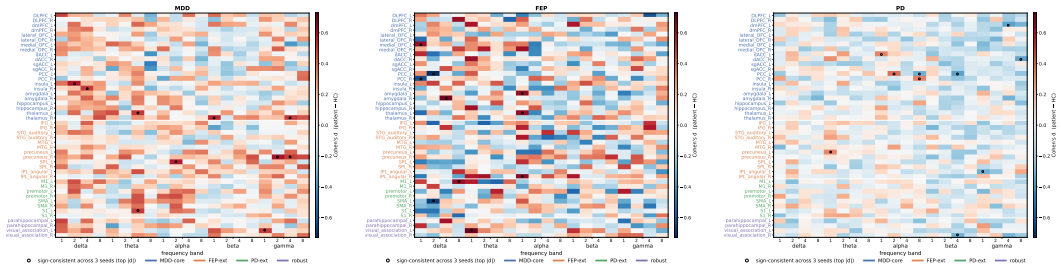
(a) MDD (K=4, r=1). (b) FEP (K=3, r=1). (c) PD (K=2, r=2).

Figure 2: **Decision-sensitivity attribution maps** of the healthy-anchored state (same 46 ROI × 5 band × 4 lag grid as Figure 3). Color is the 3-seed consensus unsigned sensitivity $\text{mean}_{\text{subj}}|\partial \,\text{logit}/\partial h_{\text{health}}|$, normalized so 1.0 is uniform (white = 0, deeper red = more sensitive); circles mark coordinates in the consensus top-10 and in $\geq 2/3$ seeds' top-20. This map asks *where the decision is sensitive*, Figure 3 *where the state differs*; the two are complementary. ROI-level attribution is cross-seed stable (axis cosine 0.83–0.87), cell-level values only weakly stable, and sensitivity differs from decision-dependence (Table 15).

J.1 ROI × FREQUENCY × LAG EFFECT MAPS

These maps read the source state along its real axes (ROI and band inherited from the input, lag from the real per-τ ACF), so each of the $46\times5\times4$ coordinates is directly readable. The decision-sensitivity attribution (Figure 2) is primary; the state-difference map (Figure 3) is a descriptive supplement.

Figure 3 shows how the state itself differs between patients and HC (mean Cohen's d of h^K over the same grid), complementing the decision-sensitivity map. The two quantify different properties and can take different patterns; their low cell-level correlation (de-meaned Pearson -0.02 to -0.11) reflects this difference.



(a) MDD (K=4, r=1). (b) FEP (K=3, r=1). (c) PD (K=2, r=2).

Figure 3: **Source-state effect maps** of the final deformed state h^K (46 ROI $\times$ 5 bands $\times$ 4 lags), same 3-seed full-cohort protocol as Figure 2. Rows are the 46 ROIs grouped by atlas origin (MDD-core 22 / FEP-ext 12 / PD-ext 8 / robust 4; white lines separate groups); columns are the five bands, each split into lags $\{1, 2, 4, 8\}$. Color is the mean Cohen's d (patient $-$ HC) of the per-subject final state on a fixed ± 1.0 axis (red $=$ patients more positive, blue $=$ more negative). Circles mark coordinates whose d sign is consistent across seeds and whose $|d|$ is in the top-k. Descriptive effect maps, distinct from significance maps.

J.2 Per-subject deformation trajectory and state space

These panels show whether a subject's deviation accumulates through iterative solving and whether the groups occupy different state-space regions and in which direction the decision pushes each subject.

J.3 Dual-view AUROC

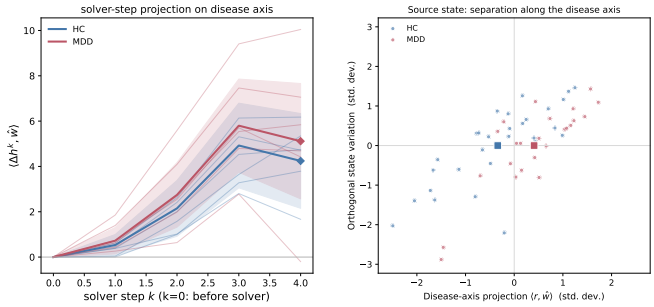
These curves compare the ranking ability of the sensor, source, and fusion on pooled holdout subjects, separating ranking (AUC) from the operating point (BAcc/F1).

J.4 Confidence scatter

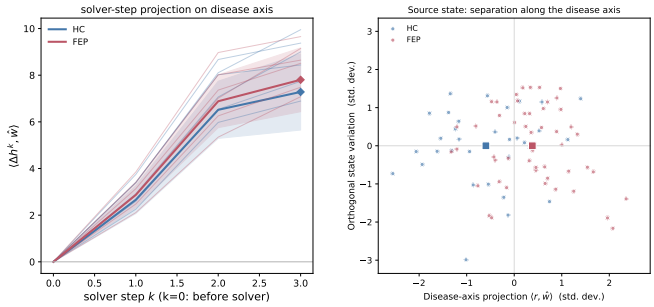
These plots show whether the branches make the same or different errors on individual subjects; different mistakes (points off the $y{=}x$ line in the disagreement quadrants) are the precondition for fusion to help.

J.5 Disease pathway glass brains

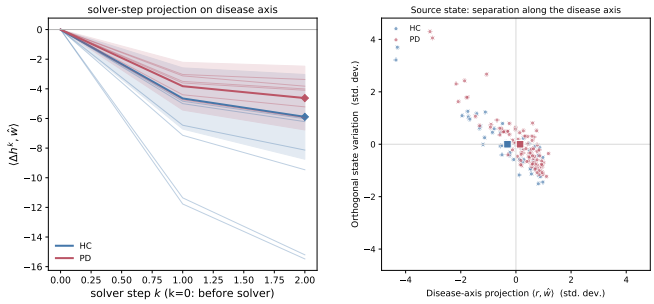
These glass brains show which source-space ROIs the decision is most sensitive to in each disease and their anatomical arrangement.



(a) MDD.

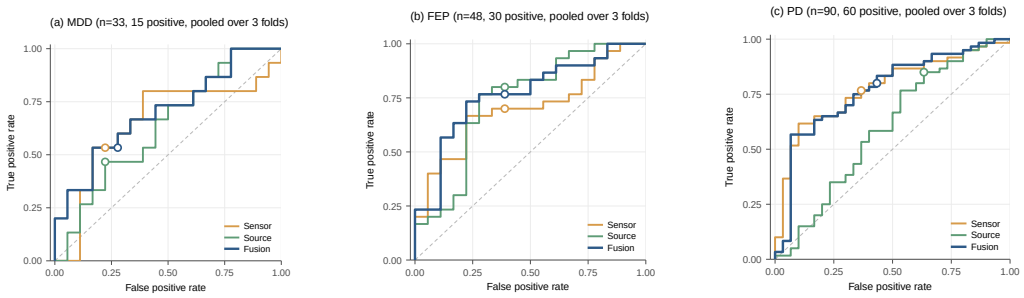


(b) FEP.



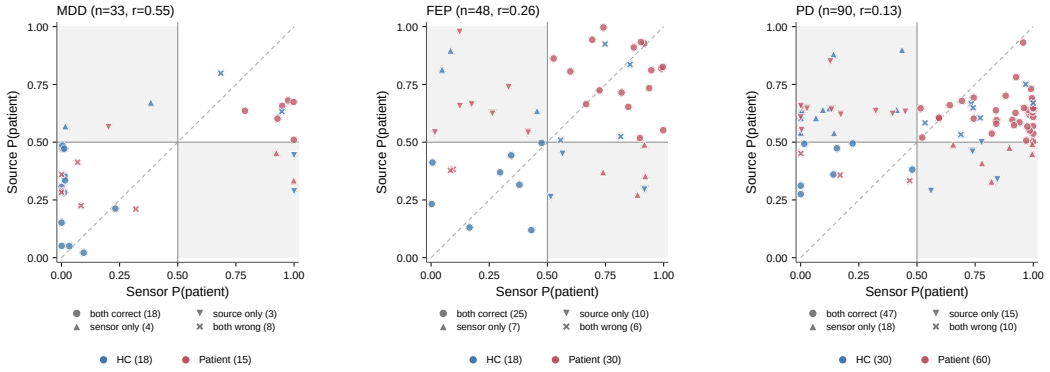
(c) PD.

Figure 4: **Per-subject deformation trajectory and state geometry** (descriptive; one representative checkpoint per panel, cv-pool subjects seen in training). Left: the disease-axis projection $\langle \Delta h^k, \hat{w} \rangle$ as the solver iterates k (thin lines, subjects; solid line, group mean $\pm$ std); monotone group separation indicates the deviation emerges during solving. Right: the two directions of largest cross-subject variance of h_{health} (920-dim) define a shared 2D plane with per-subject points (HC blue, patients red), group-mean squares, one HC$\rightarrow$patient offset arrow (direction real; length scaled for visibility, offset in σ annotated), and per-subject quiver arrows of the decision gradient $\partial \,\text{logit}/\partial h_{\text{health}}$ (direction = where the state must move to raise the patient probability; length = sensitivity). Axes are normalized per principal direction to σ units on a fixed $\pm 3.5\sigma$ range shared across diseases.



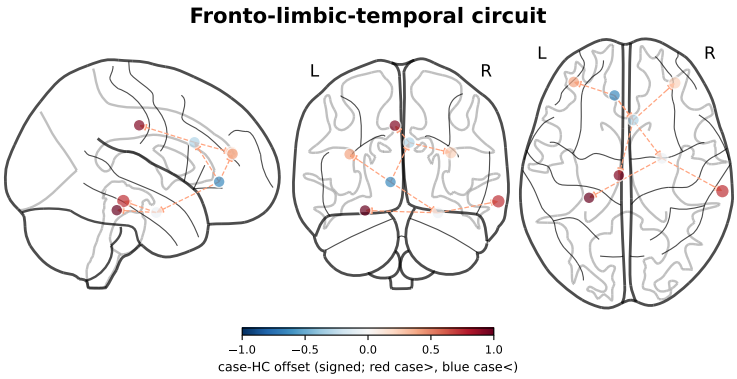
(a) MDD. (b) FEP. (c) PD.

Figure 5: **Receiver operating characteristic of the Sensor Expert, Source Expert, and DMD-EEG (Full)** on the pooled subjects of the three non-overlapping holdouts (each subject predicted once; n=33 MDD, 48 FEP, 90 PD). Fusion in dark blue, Sensor in amber, Source in sage green; open circles mark the fixed 0.5-threshold operating point that determines BAcc/F1. AUC here is the pooled value over concatenated holdout subjects and differs from the per-seed mean±std in Tables 3 and 2. Fusion is competitive with and complementary to the best single branch: on MDD a positive trend over both branches, on FEP matching the strong source branch, on PD exceeding the source branch while matching the sensor. Fusion and sensor can share the 0.5-threshold point, hence the same BAcc, while differing in AUC; the ROC separates ranking quality from the operating point. Curves are piecewise linear because the empirical ROC has finitely many thresholds.

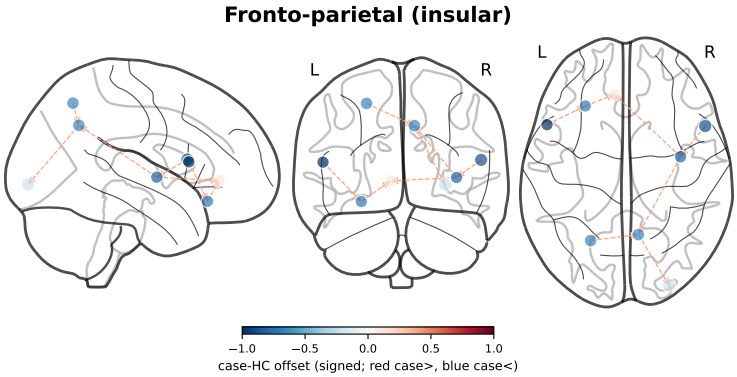


(a) MDD. (b) FEP. (c) PD.

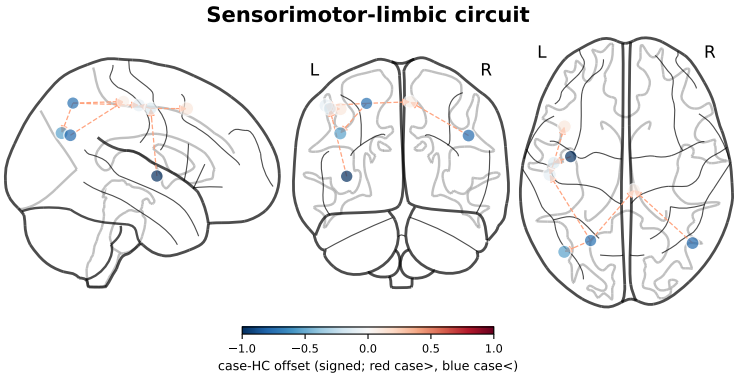
Figure 6: **Sensor versus Source predicted probabilities** (P(patient); HC blue, patients red). Shapes encode which branch is correct: ○ both, △ sensor only, ▽ source only, × both wrong (counts in legend). The 0.5 threshold lines define four quadrants; the two disagreement quadrants are shaded and the dashed y=x line marks agreement. Pooled Pearson correlations (on the panels): MDD r=0.55, FEP 0.26, PD 0.13. Points in the disagreement triangles are subjects one branch gets wrong and the other recovers; sensor-only/source-only counts are 4/3 (MDD), 7/10 (FEP), 18/15 (PD). On PD both branches err on many subjects and the source branch is weaker, so the fusion stays sensor-dominated.



(a) MDD (fronto-limbic-temporal).



(b) FEP (fronto-parietal, insular).



(c) PD (sensorimotor, insular).

Figure 7: **Disease pathway glass brains** (MNI ortho views). Two quantities are overlaid and should be kept distinct. *Node selection and size* encode decision sensitivity: the eight ROIs with the highest unsigned amplitude attribution $|\partial \,\mathrm{logit}/\partial h_{\mathrm{health}}|$ (3-seed full-cohort, cross-seed stable). *Node color* encodes the case−HC state offset $E[h \mid \mathrm{case}] - E[h \mid \mathrm{HC}]$ (red = higher in patients, blue = lower), per disease normalized; a group-level representational difference, separate from a decision quantity. The two can differ. Arrows are an illustrative minimum-spanning-tree pathway directed from smaller to larger offset, independent of learned connectivity. Circuit labels: MDD fronto-limbic-temporal, FEP fronto-parietal (insular), PD sensorimotor (insular). Colors and sizes are per-disease normalized, so absolute magnitudes are comparable only within a disease; offsets are group-level and correlational, without a causal claim, and deep ROIs (amygdala, hippocampus) are sLORETA source estimates.