BRAINFLOW: DEPRESSION DETECTION VIA TRAJECTORY DEVIATION ON GENERATIVE EEG VELOCITY FIELDS

Anonymous authors
Paper under double-blind review

ABSTRACT

Electroencephalography (EEG)-based depression detection predominantly relies on discriminative representation learning that learns decision boundaries between healthy and depressed individuals. However, substantial inter-subject variability may entangle disease-related patterns with cohort-specific variations, limiting generalization and obscuring individual deviations from healthy EEG. To address these limitations, we propose BrainFlow, a generation-based framework that learns healthy EEG velocity fields from heterogeneous datasets to establish a transferable normative reference. Yet absolute mismatch to the learned healthy dynamics can remain unreliable, since even healthy subjects may exhibit considerable residuals due to individual variability, distribution shifts, and imperfect generative modeling. We therefore introduce Trajectory Deviation, which compares unseen subjects with locally matched healthy trajectories and corrects their transport residuals using corresponding healthy references. The resulting deviations are further calibrated into physiologically interpretable Flow-Energy representations for supervised depression detection. Extensive experiments demonstrate the effectiveness of BrainFlow under subject-independent evaluation settings. (Partial code is provided in the supplementary material.)

Keywords: EEG; Depression Detection; Generative Velocity Fields; Trajectory Deviation.

1 INTRODUCTION

Electroencephalography (EEG) has emerged as a promising physiological modality for computer-aided diagnosis of major depressive disorder (MDD), owing to its non-invasive acquisition and ability to reveal functional alterations associated with depression (Guo et al., 2025; El Ouahidi et al., 2025). Recent studies have improved EEG-based depression detection through increasingly powerful discriminative representation learning, including convolutional models for local pattern extraction (Song et al., 2022; Xia et al., 2023), graph-based models for relational modeling (Sun et al., 2024; Lu et al., 2024), and attention or Transformer architectures for global dependency modeling (Ying et al., 2024; Hou et al., 2025). Despite their architectural differences, most existing approaches share essentially the same learning paradigm: depression detection is formulated as a discriminative classification problem, where models are trained to directly separate healthy control (HC) and major depressive disorder (MDD) samples.

However, this discriminative formulation faces two major limitations. First, EEG exhibits substantial inter-subject variability, causing noticeable distribution shifts even among subjects from the same dataset. As a result, generalization to unseen subjects often resembles a cross-domain generalization problem (Ng & Guan, 2024; Kaur et al., 2025) in this field. Because the decision boundary is jointly determined by the HC and MDD distributions in the training cohort, subject- or cohort-specific variations may become entangled with disease-related differences. Second, a discriminative prediction mainly indicates which side of the learned boundary a subject falls on, rather than explicitly characterizing how that subject departs from healthy EEG. These limitations motivate a different perspective: instead of learning only what separates HC from MDD, can we establish a transferable healthy reference and characterize each subject by its deviation from it?

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.

Generative models (Goodfellow et al., 2014; Ho et al., 2020) provide a natural way to establish such a healthy reference without relying on disease-specific decision boundaries. For example, models are typically trained on normal data and reduce abnormality to an endpoint-based score, such as reconstruction error, likelihood, or distance to the learned normal distribution in anomaly detection He et al. (2024); Zhang et al. (2025). Although the underlying generative model may involve an iterative process, the intermediate evolution is usually not explicitly exploited to characterize how a sample departs from normality. This can be particularly limiting for EEG, where substantial inter-subject variability may cause healthy subjects to appear atypical under a single endpoint-based score. We therefore move beyond endpoint abnormality and exploit the generative trajectory. Trajectory-based models describe a progressive transformation from a simple source distribution toward the healthy feature distribution, allowing deviations to be examined throughout the generative process rather than only at its endpoint.

Yet an absolute trajectory mismatch is still insufficient. Healthy subjects may also produce non-negligible residuals under the learned healthy dynamics because of normal individual variation and imperfect generative modeling. Therefore, a large mismatch does not necessarily imply disease-related abnormality. A more reliable measure should determine whether a subject deviates from the healthy dynamics *beyond the variation observed in comparable healthy subjects*. This motivates a relative trajectory-deviation formulation using locally matched healthy trajectories as individualized references.

To this end, we propose **BrainFlow**, a generative framework for EEG-based depression detection. BrainFlow learns healthy normative velocity fields from HC data across heterogeneous datasets to establish a shared generative reference. Based on these fields, we introduce **Velocity Trajectory Deviation Inference**, which compares the transport residuals of an unseen subject with those of locally matched healthy trajectories, thereby suppressing residual variations also present in healthy references. The resulting relative deviations are further calibrated into Flow-Energy representations that remain traceable to physiologically defined EEG features and are used for supervised MDD prediction. The deviation computation itself requires no additional training for unseen subjects. By combining multi-source healthy velocity-field modeling with reference-corrected trajectory deviation, BrainFlow provides a transferable and interpretable framework for EEG-based depression detection.

Our main contributions are summarized as follows:

- We introduce **BrainFlow**, a generative framework that learns healthy velocity fields from heterogeneous EEG datasets and reformulates depression detection as subject-specific deviation from healthy generative dynamics.
- We propose **Velocity Trajectory Deviation Inference**, which corrects subject-specific transport residuals using locally matched healthy references and constructs HC-calibrated Flow-Energy representations, providing physiologically traceable deviation features without additional training for unseen subjects.
- We conduct comprehensive subject-independent experiments across multiple public EEG datasets, together with component ablations and physiological attribution analyses, to evaluate the effectiveness and interpretability of BrainFlow.

2 Related Work

2.1 EEG-based Depression Detection

Electroencephalography (EEG)-based depression detection has been extensively studied within a discriminative learning paradigm. Traditional approaches combine handcrafted temporal, spectral, and connectivity features with conventional machine learning classifiers (Mumtaz et al., 2017). Deep learning methods further employ various architectures to learn discriminative EEG representations, including graph neural networks (GNNs) for modeling inter-channel relationships (Zhang et al., 2024), attention-based models for identifying informative EEG patterns (Ying et al., 2024), and convolutional Transformers for capturing local and long-range dependencies (Hou et al., 2025).

Despite architectural differences, existing methods mainly learn HC/MDD decision boundaries from labeled data, which may capture subject-specific patterns and obscure deviations from healthy EEG.

In contrast, BrainFlow models healthy velocity fields from HC data and quantifies individual deviations from healthy generative dynamics for subsequent MDD classification.

2.2 Generative Models for EEG Analysis

Generative models have been widely explored in EEG analysis, primarily for data augmentation and representation learning. Generative adversarial networks (GANs) and variational autoencoders (VAEs) have been employed to synthesize EEG signals and alleviate data scarcity (Hartmann et al., 2018; Luo et al., 2020). Diffusion models further enable the modeling of complex EEG distributions. For instance, DiffMDD (Wang et al., 2024b) combines diffusion-based noise-robust learning and data augmentation to improve discriminative depression detection. More recently, Flow Matching (Lipman et al., 2023) has introduced continuous-time generative modeling through learned velocity fields, with JET (Wang et al., 2026) extending this paradigm to multichannel EEG generation.

Existing generative EEG methods primarily focus on synthesizing signals or enhancing discriminative representation learning, rather than directly exploiting learned generative dynamics as diagnostic evidence. In contrast, BrainFlow models the healthy EEG distribution through a generative velocity field and introduces Trajectory Deviation to quantify the discrepancy between unseen samples and matched healthy references. By correcting subject-specific transport residuals with healthy-reference residuals, BrainFlow derives calibrated, feature-level deviation representations for supervised depression detection.

3 Methods

3.1 Overview

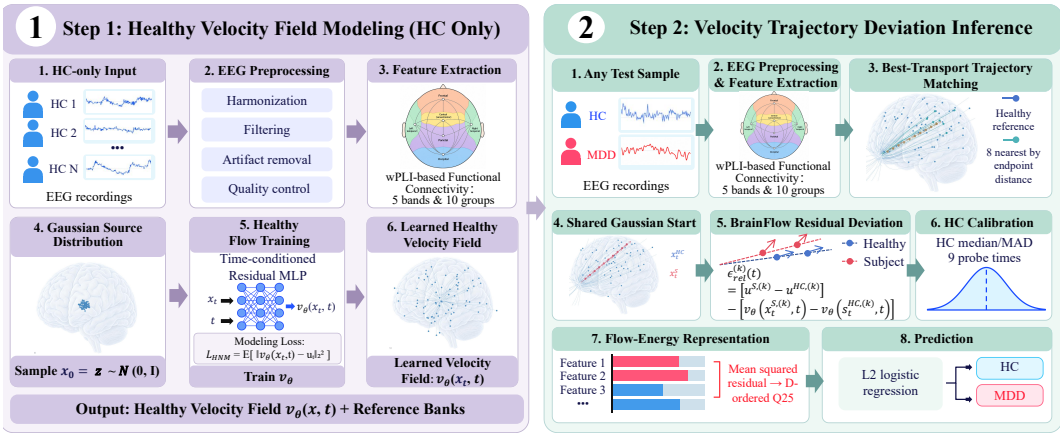

Figure 1: Overview of BrainFlow. The first stage learns a healthy velocity field and constructs healthy reference trajectories using HC only. The second stage quantifies the relative trajectory deviation of an arbitrary subject, converts it into Flow-Energy features for MDD/HC prediction.

As illustrated in Figure 1, BrainFlow consists of two stages: healthy velocity field modeling and velocity trajectory deviation inference. In the first stage, we extract hand-crafted EEG features from healthy controls (HC) and learn their generative velocity fields using generative models. The learned fields are subsequently used to construct banks of healthy reference trajectories. In the second stage, given an arbitrary subject, we retrieve locally compatible healthy trajectories and compare the subject and reference paths under the same learned velocity field. Their relative velocity residuals are calibrated using HC statistics and aggregated into a Flow-Energy representation for prediction.

3.2 Healthy Velocity Field Modeling

EEG preprocessing and feature extraction. As shown in the left panel of Figure 1, the first stage takes only HC recordings as input. We first perform channel harmonization, resampling and filtering,

artifact removal, and epoch-level quality control. From the preprocessed recordings, we extract $d_0 = 170$ subject-level handcrafted EEG features, including weighted phase-lag index (wPLI)-based regional functional connectivity and channel-wise node strength. These features are organized into ten groups across five frequency bands, with preprocessing and feature definitions detailed in Appendices B.1 and A.1.

We standardize features using normative-training HC statistics (Appendix A.2) and model each feature group with an independent velocity field. For simplicity, we omit group indices in the following formulation and denote a group-specific feature vector by $\mathbf{x} \in \mathbb{R}^{d_g}$, where d_g is the corresponding feature dimension and p_{H} denotes its HC distribution.

Healthy flow training. To model the healthy velocity fields, we adopt generative models, including flow matching (Lipman et al., 2023) and diffusion model (Ho et al., 2020), which learn a continuous transport process from a Gaussian source distribution to the target HC distribution.

Specifically, given a Gaussian initial point $\mathbf{x}_0 = \mathbf{z} \sim p_0 = \mathcal{N}(\mathbf{0}, \mathbf{I}_{d_g})$ and a healthy target sample $\mathbf{x}_1 \sim p_{\mathrm{H}}$, we construct a linear interpolation path $\mathbf{x}_t = (1-t)\mathbf{x}_0 + t\mathbf{x}_1$, where $t \in [0,1]$ is the generative transport time rather than the recording time of EEG.

For each endpoint pair, the linear interpolation defines the conditional target velocity $\mathbf{u}_t = \mathbf{x}_1 - \mathbf{x}_0$. We employ a time-conditioned residual multilayer perceptron (MLP) to learn the healthy velocity field $v_\theta(\mathbf{x}_t, t)$ by minimizing

$$\mathcal{L}_{\mathrm{HVFM}} = \mathbb{E}_{\substack{\mathbf{x}_0 \sim p_0,\, \mathbf{x}_1 \sim p_{\mathrm{H}}, \\ t \sim \mathcal{U}(0,1)}} \left[\left\| v_\theta(\mathbf{x}_t, t) - \mathbf{u}_t \right\|_2^2 \right]. \tag{1}$$

Since the target samples are drawn exclusively from HC, the learned field characterizes the generative transport dynamics of healthy EEG features without using disease labels or learning an HC/MDD decision boundary.

Healthy reference bank. After training, we use the learned velocity field to generate healthy transport trajectories by solving

$$\frac{\mathrm{d}\mathbf{x}_t^{\mathrm{flow}}}{\mathrm{d}t} = v_\theta(\mathbf{x}_t^{\mathrm{flow}}, t), \qquad \mathbf{x}_0^{\mathrm{flow}} = \mathbf{z}. \tag{2}$$

For each feature group, we sample $M = 4096$ Gaussian initial points $\{\mathbf{z}^{(k)}\}_{k=1}^{M}$ and integrate Eq. 2 from $t = 0$ to $t = 1$ to obtain the corresponding healthy trajectories and endpoints $\{\mathbf{x}_1^{\mathrm{HC},(k)}\}_{k=1}^{M}$. Each candidate is stored together with its Gaussian start as $\mathcal{T}^{(k)} = (\mathbf{z}^{(k)}, \mathbf{x}_1^{\mathrm{HC},(k)})$.

The initial points are restricted to the typical Gaussian-radius region, and ODE integration is performed using 50 explicit Euler steps. Further implementation details are provided in Appendix A.4.

The first stage therefore produces two complementary healthy references: the learned velocity field v_θ and a bank of healthy transport trajectories. Both are fixed during the subsequent deviation computation.

3.3 Velocity Trajectory Deviation Inference

Best-transport trajectory matching. Given an arbitrary test subject (HC or MDD), we apply the same preprocessing and feature extraction pipeline to obtain the corresponding groupwise feature vector $\mathbf{x}_1^{\mathrm{S}}$. We then select locally compatible healthy trajectories from the reference bank according to their endpoint distances.

Before matching, we apply HC-only normative quality control (HCQC-A) to retain feature coordinates that are reliably represented by the learned healthy reference. For each coordinate, HCQC-A evaluates healthy-reference fidelity using location agreement, scale agreement, generated-reference statistics, and extreme-score stability, and retains features with normative-quality score $Q_{\mathrm{HC},j} \geq 0.80$. Importantly, this screening uses HC observations only; MDD and test participants are not involved. Let $\mathcal{J}_g$ denote the retained feature coordinates in group g. The compatibility between the subject and the k-th healthy reference is measured by

$$D^{(k)}(\mathbf{x}_1^{\mathrm{S}}) = \frac{1}{|\mathcal{J}_g|} \sum_{j \in \mathcal{J}_g} \left(x_{1,j}^{\mathrm{S}} - x_{1,j}^{\mathrm{HC},(k)} \right)^2. \tag{3}$$

We select the $K = 8$ trajectories with the smallest endpoint distances and arrange them in ascending order of $D^{(k)}$. Matching is performed over the HCQC-A-retained coordinates, while trajectory generation and velocity-field evaluation operate in the full group feature space. The complete definition of HCQC-A is provided in Appendix A.5.

BrainFlow residual deviation. For each matched reference, we construct two linear transport paths sharing the same Gaussian initial point $\mathbf{z}^{(k)}$:

$$\begin{aligned} \mathbf{x}_t^{\mathrm{S},(k)} &= (1-t)\mathbf{z}^{(k)} + t\mathbf{x}_1^{\mathrm{S}}, \\ \mathbf{x}_t^{\mathrm{HC},(k)} &= (1-t)\mathbf{z}^{(k)} + t\mathbf{x}_1^{\mathrm{HC},(k)}. \end{aligned} \tag{4}$$

These linear comparison paths are distinct from the generative ODE trajectories used to construct the healthy reference bank. Specifically, the reference-bank trajectories are obtained by integrating the learned healthy velocity field from Gaussian initial points, whereas the linear paths above are introduced only for paired subject-reference comparison. Their target velocities are $\mathbf{u}^{\mathrm{S},(k)} = \mathbf{x}_1^{\mathrm{S}} - \mathbf{z}^{(k)}$ and $\mathbf{u}^{\mathrm{HC},(k)} = \mathbf{x}_1^{\mathrm{HC},(k)} - \mathbf{z}^{(k)}$.

We evaluate both paths under the same learned healthy velocity field and define their respective transport residuals as

$$\begin{aligned} \mathbf{r}^{\mathrm{S},(k)}(t) &= \mathbf{u}^{\mathrm{S},(k)} - v_\theta(\mathbf{x}_t^{\mathrm{S},(k)}, t), \\ \mathbf{r}^{\mathrm{HC},(k)}(t) &= \mathbf{u}^{\mathrm{HC},(k)} - v_\theta(\mathbf{x}_t^{\mathrm{HC},(k)}, t). \end{aligned} \tag{5}$$

The BrainFlow residual is then defined as their relative difference:

$$\begin{aligned} \boldsymbol{\epsilon}_{\mathrm{rel}}^{(k)}(t) &= \mathbf{r}^{\mathrm{S},(k)}(t) - \mathbf{r}^{\mathrm{HC},(k)}(t) \\ &= \left[\mathbf{u}^{\mathrm{S},(k)} - \mathbf{u}^{\mathrm{HC},(k)}\right] - \left[v_\theta(\mathbf{x}_t^{\mathrm{S},(k)}, t) - v_\theta(\mathbf{x}_t^{\mathrm{HC},(k)}, t)\right]. \end{aligned} \tag{6}$$

Since both paths share the same Gaussian start, their target velocity difference equals the displacement between the subject and healthy-reference endpoints. The relative residual thus measures the difference in transport behavior after correcting for the healthy velocity-field difference along the paired paths, rather than relying on endpoint distance alone.

HC calibration and Flow-Energy representation. To account for variations across feature coordinates and transport times, we calibrate the relative residuals using HC statistics.

Specifically, we evaluate $\boldsymbol{\epsilon}_{\mathrm{rel}}^{(k)}(t)$ at nine uniformly spaced probe times t_m between 0.05 and 0.95. For each feature coordinate j, the residual is standardized as

$$\widetilde{\epsilon}_{\mathrm{rel},j}^{(k)}(t_m) = \frac{\epsilon_{\mathrm{rel},j}^{(k)}(t_m) - \mathrm{median}_{\mathrm{HC},j}(t_m)}{s_{\mathrm{HC},j}(t_m)}, \tag{7}$$

where $s_{\mathrm{HC},j}(t_m) = 1.4826\ \mathrm{MAD}_{\mathrm{HC},j}(t_m)$, where MAD denotes the median absolute deviation. If the robust scale is non-finite or below 10^{-4}, it is replaced by the corresponding dataset-macro-weighted standard deviation and lower-bounded by 10^{-4}. The calibration statistics are estimated exclusively from source HC, excluding MDD and held-out test subjects (Appendix A.7).

We then calculate the mean squared standardized residual along each matched path and aggregate the resulting energies using distance-ordered interpolation:

$$E_j(\mathbf{x}_1^{\mathrm{S}}) = Q^D_{0.25,\, k\in\mathcal{K}(\mathbf{x}_1^{\mathrm{S}})}\left[\frac{1}{9}\sum_{m=1}^{9}\left(\widetilde{\epsilon}_{\mathrm{rel},j}^{(k)}(t_m)\right)^2\right]. \tag{8}$$

Here, $Q^D_{0.25}$ denotes linear interpolation at the 25% position after ordering the eight paths by endpoint distance rather than by energy. For $K = 8$, this combines the second- and third-nearest path energies with weights 0.25 and 0.75, respectively. The resulting E_j quantifies the feature-specific deviation from locally matched healthy transport behavior.

To capture complementary levels of deviation, we concatenate detail-level Flow-Energy features with group-level and global-level scores:

$$\mathbf{E}_{\mathrm{all}} = \left[\mathbf{E}_{\mathrm{detail}}^\top, \mathbf{E}_{\mathrm{group}}^\top, E_{\mathrm{global}}\right]^\top \in \mathbb{R}^{d+G+1}. \tag{9}$$

Here, $\mathbf{E}_{\text{detail}} = [E_1, \ldots, E_d]^\top$ contains the retained feature-level energies, where $d \leq d_0$. The vector $\mathbf{E}_{\text{group}} \in \mathbb{R}^G$ contains one joint deviation score for each retained, nonempty feature group ($G \leq 10$), while E_{global} summarizes the group-level deviations. These joint scores are constructed using shrinkage-regularized Mahalanobis distances and HC empirical tail calibration, as detailed in Appendix A.7.

Prediction and physiological attribution. Finally, we apply a logarithmic transformation and training-set standardization to obtain

$$\mathbf{h} = \mathcal{S}_{\text{train}} \left(\log(1 + \mathbf{E}_{\text{all}}) \right), \tag{10}$$

where $\mathcal{S}_{\text{train}}$ performs median/MAD standardization and source-HC site-drift attenuation using training data only.

An L_2-regularized logistic regression classifier fitted on labeled HC and MDD training subjects estimates

$$P(y = \text{MDD} \mid \mathbf{x}_1^{\text{S}}) = \sigma \left(\mathbf{w}^\top \mathbf{h} + b \right). \tag{11}$$

Once the healthy normative field, reference bank, calibration statistics, and prediction classifier have been established, inference on a new subject requires no parameter updates. Moreover, each detail-level Flow-Energy coordinate can be traced back to an original EEG descriptor, while its corresponding contribution $c_j = w_j h_j$ quantifies its signed contribution to the prediction logit. This enables feature-level attribution in terms of EEG frequency bands, channel-wise node strengths, and regional wPLI connections.

4 Experiments

4.1 Experimental Settings

Datasets. As summarized in Figure 2, the experiments are conducted on five public EEG datasets: Dortmund (Wascher et al., 2025), LEMON (Babayan et al., 2019), MODMA_128 (Cai et al., 2022), Mumtaz (Mumtaz, 2016), and TDBRAIN (van Dijk et al., 2022). The resulting cohort contains 942 participants with 35,271 non-overlapping 5-s EEG segments after quality control.

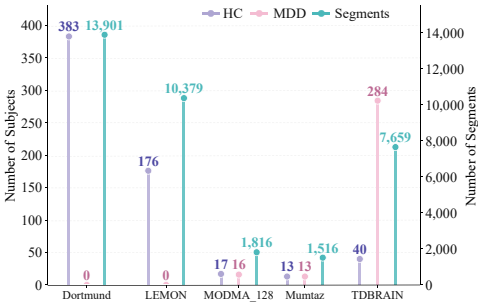

Figure 2: Dataset statistics.

To ensure strict separation between the two stages of BrainFlow, all participants are partitioned at the subject level before model development. Step 1 uses 472 HC participants, comprising 101,125 seconds of EEG and 20,225 segments, to learn the healthy velocity fields and construct the reference banks. Step 2 uses a disjoint set of 157 HC and 313 MDD participants, comprising 6,727 and 8,319 segments, respectively. Thus, no subject or EEG segment used to learn the healthy normative flow is reused in Step 2. Test subjects are further excluded from all calibration and classifier-training procedures.

Implementation Details. Each feature group is modeled by a time-conditioned residual MLP with a 64-dimensional time embedding, 256 hidden units, three residual blocks, and a dropout rate of 0.10. We train each model for 200 epochs using AdamW with a batch size of 64, a learning rate of 3×10^{-4}, and a weight decay of 10^{-5}. We evaluate performance using precision (Pre), recall (Rec), accuracy (Acc), and F1-score (F1). More implementation details and parameter validation results are provided in the Appendix.

4.2 Main Results

To ensure a fair comparison, we reproduce all competing methods under an identical subject-level data-access protocol. Specifically, each method has access to the same non-test participants, including the 472 HC subjects used by BrainFlow for healthy normative modeling; no additional training subjects are introduced for BrainFlow. The same test participants and evaluation metrics are used

across all methods. They cover complementary methodological families, including conventional deep learning models, flow-based methods, one-class approaches, and EEG foundation models, providing a broad comparison. Comparison protocols and configurations are listed in Appendix B.5.

As shown in Table 1, BrainFlow demonstrates stronger overall performance across these method families. Compared with conventional deep learning models, it avoids relying solely on a dataset-specific discriminative boundary by explicitly modeling deviations from healthy EEG. It also consistently improves over standard flow-based and one-class approaches, indicating that simply modeling the healthy distribution or measuring uncorrected abnormality is insufficient, while the proposed healthy-reference-corrected trajectory deviation provides more informative diagnostic representations. BrainFlow further remains competitive against pretrained EEG foundation models despite using a substantially simpler task-specific framework. The consistent performance of different BrainFlow variants also supports the generality of the proposed trajectory-deviation methods.

Table 1: Comparison results under the fixed dataset-role split. Values are mean ± standard deviation over five model seeds. Bold and underlined values indicate the highest and second-highest means in each metric, respectively. Parentheses indicate the generative model used.

Family	Method	Classification metrics			
		Pre ↑	Rec ↑	Acc ↑	F1 ↑
Deep learning	GNN Scarselli et al. (2009)	0.664 ± 0.016	0.664 ± 0.027	0.562 ± 0.021	0.646 ± 0.019
	LSTM Hochreiter & Schmidhuber (1997)	0.707 ± 0.022	0.798 ± 0.023	0.606 ± 0.030	0.733 ± 0.018
	Transformer Vaswani et al. (2017)	0.714 ± 0.010	**0.848 ± 0.013**	0.650 ± 0.013	0.764 ± 0.009
Flow-based Learning	Standard FM Lipman et al. (2023)	0.706 ± 0.016	0.781 ± 0.016	0.621 ± 0.017	0.731 ± 0.016
	Flow Mismatching Chen et al. (2026)	0.685 ± 0.016	0.480 ± 0.042	0.523 ± 0.025	0.559 ± 0.028
	TCCM Li et al. (2025)	<u>0.901 ± 0.003</u>	0.623 ± 0.006	0.657 ± 0.006	0.712 ± 0.006
One-class Learning	Shrinkage Mahalanobis Ledoit & Wolf (2004)	0.883 ± 0.000	0.388 ± 0.000	0.491 ± 0.000	0.501 ± 0.000
	One-Class SVM Schölkopf et al. (2001)	**0.913 ± 0.000**	0.406 ± 0.000	0.529 ± 0.000	0.545 ± 0.000
	Isolation Forest Liu et al. (2008)	0.901 ± 0.005	0.395 ± 0.003	0.517 ± 0.003	0.529 ± 0.005
EEG Model	EEGPT Wang et al. (2024a)	0.836 ± 0.027	0.695 ± 0.013	0.680 ± 0.032	0.751 ± 0.019
	CBraMod Wang et al. (2025)	0.761 ± 0.053	0.652 ± 0.005	0.609 ± 0.044	0.682 ± 0.022
	ST-EEGFormer-small Yang et al. (2026)	0.808 ± 0.018	0.755 ± 0.022	<u>0.680 ± 0.023</u>	<u>0.766 ± 0.017</u>
Proposed	**BrainFlow** (Diffusion Model)	0.727 ± 0.021	0.813 ± 0.068	0.643 ± 0.040	0.755 ± 0.037
	BrainFlow (Flow Matching)	0.753 ± 0.025	<u>0.844 ± 0.036</u>	**0.686 ± 0.026**	**0.786 ± 0.019**

4.3 Ablation Studies

Progressive component ablation. As illustrated in Figure 3, we progressively introduce healthy velocity field modeling (HVFM) and velocity trajectory deviation inference (VTDI) under the same evaluation setting. HVFM alone brings only modest and metric-dependent changes over the Base model, suggesting that healthy velocity field modeling provides useful information but is insufficient for robust HC/MDD discrimination. This is likely because absolute transport residuals still contain variations caused by inter-subject differences. Incorporating VTDI consistently improves all evaluation metrics by comparing each subject with matched healthy references and suppressing residual variations also observed in HC. These results indicate that HVFM establishes a useful normative foundation, while VTDI more effectively converts it into discriminative subject-specific deviation features.

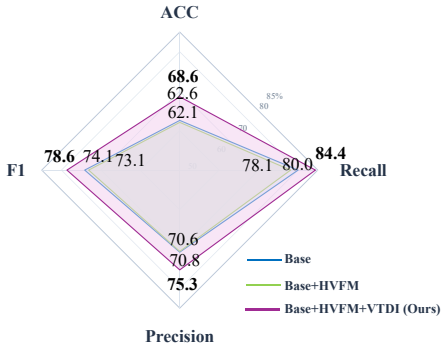

Figure 3: Progressive component ablation under the fixed dataset-role split.

Component replacement ablations. To further validate the proposed design, we replace the deviation formulation and trajectory-selection strategy while keeping the remaining pipeline unchanged.

Table 2: Component replacement ablations. Bold values indicate the highest mean in each metric.

Replaced component	Variant	Pre ↑	Rec ↑	Acc ↑	F1 ↑
Deviation feature	Absolute residual	0.732 ± 0.029	0.794 ± 0.097	0.640 ± 0.049	0.748 ± 0.053
	Generated-HC CDF	0.733 ± 0.021	0.705 ± 0.024	0.619 ± 0.027	0.713 ± 0.022
	Matched endpoint distance	0.720 ± 0.042	0.806 ± 0.082	0.637 ± 0.065	0.749 ± 0.057
Trajectory selection	Random-8 Trajectories	0.729 ± 0.025	0.806 ± 0.025	0.639 ± 0.026	0.755 ± 0.016
BrainFlow (Flow Matching)	-	$\mathbf{0.753 \pm 0.025}$	$\mathbf{0.844 \pm 0.036}$	$\mathbf{0.686 \pm 0.026}$	$\mathbf{0.786 \pm 0.019}$

As shown in Table 2, Absolute residual and Generated-HC cumulative distribution function (CDF) underperform the proposed relative deviation, confirming the benefit of healthy-reference correction. Notably, matched endpoint distance uses the same nearest healthy references as BrainFlow but removes velocity-field evaluation and measures deviation only from endpoint differences. Its inferior performance shows that local reference matching alone is insufficient and that the learned transport dynamics provide additional diagnostic information. Randomizing the healthy-reference trajectories further degrades performance, demonstrating the importance of locally compatible matching. Together, these results validate the complementary roles of healthy-reference correction, trajectory-aware deviation modeling, and subject-specific reference selection.

4.4 Physiological Attribution Analysis

To examine the interpretability of BrainFlow, we visualize how raw EEG features are transformed into healthy-reference deviations and ultimately contribute to HC/MDD prediction, as illustrated in Figure 4. Nodes represent channel-wise node strengths, while edges represent regional wPLI-based functional connectivity. Figure 4(a) presents the original connectivity patterns, which alone do not indicate whether a feature is abnormal or diagnostically relevant. In contrast, Figure 4(b) visualizes the corresponding Flow-Energy representations, highlighting feature-specific deviations from the learned healthy reference. Finally, Figure 4(c) maps these deviations to their signed contributions to the classifier, distinguishing features that support MDD prediction from those favoring HC. Together, these visualizations illustrate how BrainFlow transforms EEG connectivity descriptors into physiologically traceable deviation features and identifies their contributions to individual predictions.

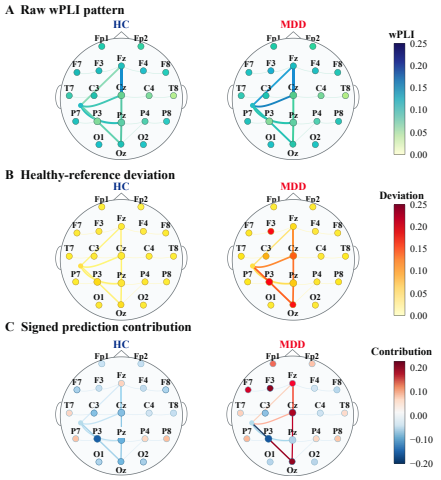

Figure 4: Raw wPLI features, healthy-reference deviations, and signed prediction contributions derived from held-out HC and MDD participants.

4.5 Spatial and Spectral Analysis of Healthy-Reference Deviations

To assess whether BrainFlow captures physiologically meaningful deviations, we analyze their spatial and spectral distributions for test subjects. A feature is considered deviating when its Flow-Energy exceeds the feature-specific 95th percentile estimated from 135 calibration HC. As shown in Figure 5, MDD and HC exhibit distinct patterns across channels, regions, and frequency bands. MDD shows more frequent node-strength deviations at F3 and Cz, greater regional deviations in central and parietal areas, and higher theta-band deviation prevalence, whereas HC shows more occipital and delta-band deviations. These results suggest that BrainFlow captures physiologically interpretable spatial and spectral differences in healthy-reference deviations.

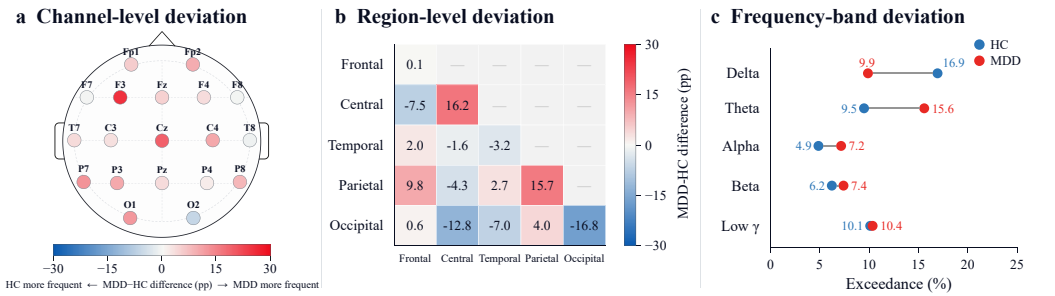

Figure 5: Spatial and spectral patterns of healthy-reference deviations in HC and MDD. Differences in deviation prevalence are visualized at the channel, regional, and frequency-band levels.

4.6 Prediction Contribution Analysis

To examine how BrainFlow's deviation features affect prediction, we analyze signed feature contributions and decompose the classifier logit across representation levels. As shown in Figure 6, influential features span node strengths and regional connectivity across multiple frequency bands, with different contribution directions for HC and MDD. Detail-level features account for most of the HC/MDD logit difference, while group- and global-level contributions are smaller. These results show that BrainFlow provides a traceable link from healthy-reference deviations to physiologically defined EEG features and final predictions. Feature-level statistics and an individual example are provided in Appendix C.3.

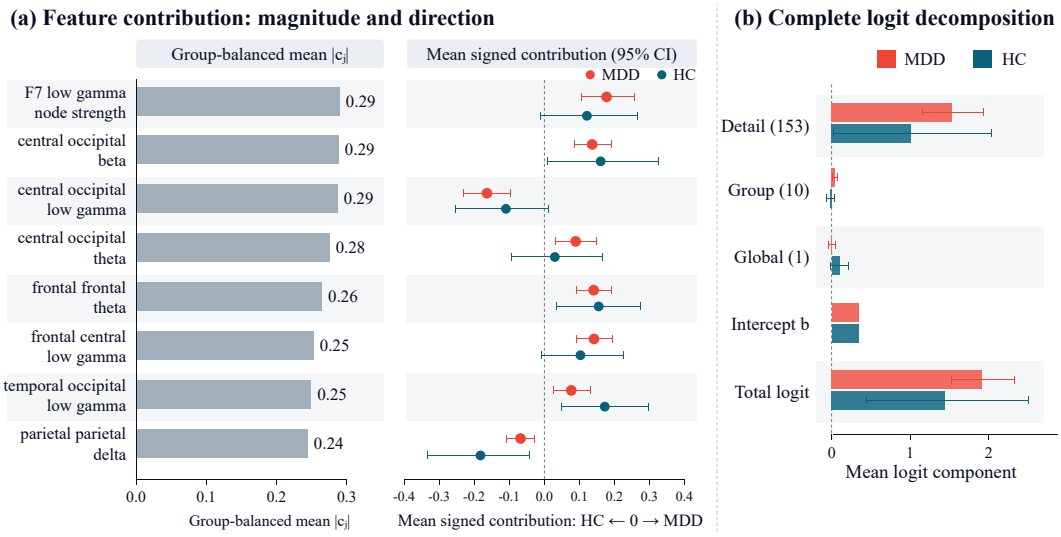

Figure 6: Feature-level prediction contributions and logit decomposition in BrainFlow. CI denotes confidence interval.

5 Conclusion

We present BrainFlow, a generative normative framework for EEG-based depression detection that learns healthy velocity fields from heterogeneous datasets. By introducing Trajectory Deviation, BrainFlow measures subject-specific deviations relative to matched healthy trajectories and constructs physiologically interpretable Flow-Energy representations. Extensive subject-independent experiments demonstrate competitive detection performance, while ablation and attribution analyses highlight the effectiveness and physiological traceability of healthy-reference trajectory deviation.

AI USE STATEMENT

Generative AI tools assisted with language polishing and literature and information retrieval. The authors verified the retrieved information and references against their original sources and take full responsibility for the final content of this paper.

ETHICS STATEMENT

This study analyzes previously released EEG datasets. We report group-level results and individual examples without identifiable personal information. The proposed model is intended for research on the auxiliary identification of MDD and is not a standalone clinical diagnostic tool.

REPRODUCIBILITY STATEMENT

Detailed implementation of BrainFlow, including EEG feature construction, healthy velocity field modeling, HC-only quality control, trajectory matching, and Flow-Energy computation, is provided in Appendix A. Dataset preprocessing, evaluation protocols, and baseline and ablation configurations are described in Appendix B. Additional feature-retention, sensitivity, and attribution analyses are reported in Appendix C.

REFERENCES

Anahit Babayan, Miray Erbey, Deniz Kumral, Janis D. Reinelt, Andrea M. F. Reiter, Josefin Röbbig, H. Lina Schaare, Marie Uhlig, Alfred Anwander, Pierre-Louis Bazin, Annette Horstmann, Leonie Lampe, Vadim V. Nikulin, Hadas Okon-Singer, Sven Preusser, André Pampel, Christiane S. Rohr, Julia Sacher, Angelika Thöne-Otto, Sabrina Trapp, Till Nierhaus, Denise Altmann, Katrin Arelin, Maria Blöchl, Edith Bongartz, Patric Breig, Elena Cesnaite, Sufang Chen, Roberto Cozatl, Saskia Czerwonatis, Gabriele Dambrauskaite, Maria Dreyer, Jessica Enders, Melina Engelhardt, Marie Michele Fischer, Norman Forschack, Johannes Golchert, Laura Golz, C. Alexandrina Guran, Susanna Hedrich, Nicole Hentschel, Daria I. Hoffmann, Julia M. Huntenburg, Rebecca Jost, Anna Kosatschek, Stella Kunzendorf, Hannah Lammers, Mark E. Lauckner, Keyvan Mahjoory, Ahmad S. Kanaan, Natacha Mendes, Ramona Menger, Enzo Morino, Karina Näthe, Jennifer Neubauer, Handan Noyan, Sabine Oligschläger, Patricia Panczyszyn-Trzewik, Dorothee Poehlchen, Nadine Putzke, Sabrina Roski, Marie-Catherine Schaller, Anja Schieferbein, Benito Schlaak, Robert Schmidt, Krzysztof J. Gorgolewski, Hanna Maria Schmidt, Anne Schrimpf, Sylvia Stasch, Maria Voss, Annett Wiedemann, Daniel S. Margulies, Michael Gaebler, and Arno Villringer. A mind-brain-body dataset of MRI, EEG, cognition, emotion, and peripheral physiology in young and old adults. *Scientific Data*, 6:180308, 2019. doi: 10.1038/sdata.2018.308.

Nima Bigdely-Shamlo, Tim Mullen, Christian Kothe, Kyung-Min Su, and Kay A. Robbins. The PREP pipeline: standardized preprocessing for large-scale EEG analysis. *Frontiers in Neuroinformatics*, 9:16, 2015. doi: 10.3389/fninf.2015.00016. URL https://doi.org/10.3389/fninf.2015.00016.

Hanshu Cai, Zhenqin Yuan, Yiwen Gao, Shuting Sun, Na Li, Fuze Tian, Han Xiao, Jianxiu Li, Zhengwu Yang, Xiaowei Li, Qinglin Zhao, Zhenyu Liu, Zhijun Yao, Minqiang Yang, Hong Peng, Jing Zhu, Xiaowei Zhang, Guoping Gao, Fang Zheng, Rui Li, Zhihua Guo, Rong Ma, Jing Yang, Lan Zhang, Xiping Hu, Yumin Li, and Bin Hu. A multi-modal open dataset for mental-disorder analysis. *Scientific Data*, 9:178, 2022. doi: 10.1038/s41597-022-01211-x.

Shengzhe Chen, Mehrdad Moradi, Kamran Paynabar, and Hao Yan. Flow mismatching: Unsupervised anomaly detection via velocity discrepancies in flow matching models. arXiv preprint arXiv:2605.23070, 2026. URL https://arxiv.org/abs/2605.23070.

Yassine El Ouahidi, Jonathan Lys, Philipp Thölke, Nicolas Farrugia, Bastien Pasdeloup, Vincent Gripon, Karim Jerbi, and Giulia Lioi. REVE: A foundation model for EEG – adapting to any setup with large-scale pretraining on 25,000 subjects. In *Advances in Neural Information Processing Systems*, volume 38, pp. 25583–25619, 2025. doi: 10.52202/

085713-0760. URL https://papers.neurips.cc/paper_files/paper/2025/hash/20a917f77773ac0fa8bea2bdd6606b66-Abstract-Conference.html.

Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, volume 27, 2014.

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

Hanning Guo, Farah Abdellatif, Yu Fu, N. Jon Shah, Abigail Morrison, and Jürgen Dammers. XAIguiFormer: explainable artificial intelligence guided transformer for brain disorder identification. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=AD5yx2xq8R.

Kay Gregor Hartmann, Robin Tibor Schirrmeister, and Tonio Ball. EEG-GAN: Generative adversarial networks for electroencephalograhic (EEG) brain signals. arXiv preprint arXiv:1806.01875, 2018. URL https://arxiv.org/abs/1806.01875.

Haoyang He, Jiangning Zhang, Hongxu Chen, Xuhai Chen, Zhishan Li, Xu Chen, Yabiao Wang, Chengjie Wang, and Lei Xie. A diffusion-based framework for multi-class anomaly detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 8472–8480, 2024. doi: 10.1609/aaai.v38i8.28690.

Jonathan Ho, Ajay N. Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, pp. 6840–6851, 2020.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8): 1735–1780, 1997. doi: 10.1162/neco.1997.9.8.1735.

Pengfei Hou, Xiaowei Li, Jing Zhu, and Bin Hu. A lightweight convolutional transformer neural network for EEG-based depression recognition. *Biomedical Signal Processing and Control*, 100: 107112, 2025. doi: 10.1016/j.bspc.2024.107112.

Mainak Jas, Denis A. Engemann, Yousra Bekhti, Federico Raimondo, and Alexandre Gramfort. Autoreject: Automated artifact rejection for MEG and EEG data. *NeuroImage*, 159:417–429, 2017. doi: 10.1016/j.neuroimage.2017.06.030. URL https://doi.org/10.1016/j.neuroimage.2017.06.030.

Gagandeep Kaur, Himanshu Aggarwal, and Neelam Goel. Artificial intelligence driven neuropsychiatry: A systematic review of electroencephalography-based computational techniques for major depressive disorder prediction. *Neuroscience*, 581:179–207, 2025. doi: 10.1016/j.neuroscience.2025.07.010.

Olivier Ledoit and Michael Wolf. A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis*, 88(2):365–411, 2004. doi: 10.1016/S0047-259X(03)00096-4.

Zhong Li, Qi Huang, Yuxuan Zhu, Lincen Yang, Mohammad Mohammadi Amiri, Niki van Stein, and Matthijs van Leeuwen. Scalable, explainable and provably robust anomaly detection with one-step flow matching. In *Advances in Neural Information Processing Systems*, volume 38, pp. 97620–97685, 2025. doi: 10.52202/085713-2936.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=PqvMRDCJT9t.

Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. Isolation forest. In *2008 Eighth IEEE International Conference on Data Mining*, pp. 413–422, 2008. doi: 10.1109/ICDM.2008.17.

Haifeng Lu, Zhiyang You, Yi Guo, and Xiping Hu. MAST-GCN: Multi-scale adaptive spatial-temporal graph convolutional network for EEG-based depression recognition. *IEEE Transactions on Affective Computing*, 15(4):1985–1996, 2024. doi: 10.1109/TAFFC.2024.3392904.

Yun Luo, Li-Zhen Zhu, Zi-Yu Wan, and Bao-Liang Lu. Data augmentation for enhancing EEG-based emotion recognition with deep generative models. *Journal of Neural Engineering*, 17(5): 056021, 2020. doi: 10.1088/1741-2552/abb580.

Wajid Mumtaz. MDD patients and healthy controls EEG data (new). figshare, 2016. URL https://doi.org/10.6084/m9.figshare.4244171.v2. Version 2.

Wajid Mumtaz, Likun Xia, Syed Saad Azhar Ali, Mohd Azhar Mohd Yasin, Muhammad Hussain, and Aamir Saeed Malik. Electroencephalogram (EEG)-based computer-aided technique to diagnose major depressive disorder (MDD). *Biomedical Signal Processing and Control*, 31:108–115, 2017. doi: 10.1016/j.bspc.2016.07.006.

Han Wei Ng and Cuntai Guan. Subject-independent meta-learning framework towards optimal training of EEG-based classifiers. *Neural Networks*, 172:106108, 2024. doi: 10.1016/j.neunet.2024.106108.

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

Franco Scarselli, Marco Gori, Ah Chung Tsoi, Markus Hagenbuchner, and Gabriele Monfardini. The graph neural network model. *IEEE Transactions on Neural Networks*, 20(1):61–80, 2009. doi: 10.1109/TNN.2008.2005605.

Bernhard Schölkopf, John C. Platt, John Shawe-Taylor, Alex J. Smola, and Robert C. Williamson. Estimating the support of a high-dimensional distribution. *Neural Computation*, 13(7):1443–1471, 2001. doi: 10.1162/089976601750264965.

XinWang Song, DanDan Yan, LuLu Zhao, and LiCai Yang. LSDD-EEGNet: An efficient end-to-end framework for EEG-based depression detection. *Biomedical Signal Processing and Control*, 75:103612, 2022. doi: 10.1016/j.bspc.2022.103612.

Xiaofang Sun, Yonghui Xu, Yibowen Zhao, Xiangwei Zheng, Yongqing Zheng, and Lizhen Cui. Multi-granularity graph convolution network for major depressive disorder recognition. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 32:559–569, 2024. doi: 10.1109/TNSRE.2023.3311458.

Hanneke van Dijk, Guido van Wingen, Damiaan Denys, Sebastian Olbrich, Rosalinde van Ruth, and Martijn Arns. The two decades brainclinics research archive for insights in neurophysiology (TDBRAIN) database. *Scientific Data*, 9:333, 2022. doi: 10.1038/s41597-022-01409-z.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, pp. 5998–6008, 2017. URL https://papers.nips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

Martin Vinck, Robert Oostenveld, Marijn van Wingerden, F. Battaglia, and Cyriel M. A. Pennartz. An improved index of phase-synchronization for electrophysiological data in the presence of volume-conduction, noise and sample-size bias. *NeuroImage*, 55(4):1548–1565, 2011. doi: 10.1016/j.neuroimage.2011.01.055. URL https://doi.org/10.1016/j.neuroimage.2011.01.055.

Guagnyu Wang, Wenchao Liu, Yuhong He, Cong Xu, Lin Ma, and Haifeng Li. EEGPT: Pretrained transformer for universal and reliable representation of EEG signals. In *Advances in Neural Information Processing Systems*, volume 37, pp. 39249–39280, 2024a. doi: 10.52202/079017-1239.

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*, 2025. URL https://openreview.net/forum?id=NPNUHgHF2w.

Yifan Wang, Yijia Ma, Wen Li, and Chenyu You. Let EEG models learn EEG. In *International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2605.21280.

Yilin Wang, Sha Zhao, Haiteng Jiang, Shijian Li, Benyan Luo, Tao Li, and Gang Pan. DiffMDD: A diffusion-based deep learning framework for MDD diagnosis using EEG. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 32:728–738, 2024b. doi: 10.1109/TNSRE.2024.3360465.

Edmund Wascher, Daniel Schneider, Patrick D. Gajewski, and Stephan Getzmann. Resting-state EEG data before and after cognitive activity across the adult lifespan and a 5-year follow-up. OpenNeuro dataset, version 1.0.3, 2025. URL https://doi.org/10.18112/openneuro.ds005385.v1.0.3.

Min Xia, Yangsong Zhang, Yihan Wu, and Xiuzhu Wang. An end-to-end deep learning model for EEG-based major depressive disorder classification. *IEEE Access*, 11:41337–41347, 2023. doi: 10.1109/ACCESS.2023.3270426.

Liuyin Yang, Qiang Sun, Ang Li, and Marc M. Van Hulle. Are EEG foundation models worth it? comparative evaluation with traditional decoders in diverse BCI tasks. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=5Xwm8e6vbh.

Ming Ying, Xuexiao Shao, Jing Zhu, Qinglin Zhao, Xiaowei Li, and Bin Hu. EDT: An EEG-based attention model for feature learning and depression recognition. *Biomedical Signal Processing and Control*, 93:106182, 2024. doi: 10.1016/j.bspc.2024.106182.

Hui Zhang, Zheng Wang, Dan Zeng, Zuxuan Wu, and Yu-Gang Jiang. DiffusionAD: Norm-guided one-step denoising diffusion for anomaly detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(8):7140–7152, 2025. doi: 10.1109/TPAMI.2025.3570494.

Zhongyi Zhang, Qinghao Meng, LiCheng Jin, Hanguang Wang, and Huirang Hou. A novel EEG-based graph convolution network for depression detection: Incorporating secondary subject partitioning and attention mechanism. *Expert Systems with Applications*, 239:122356, 2024. doi: 10.1016/j.eswa.2023.122356.

A Implementation Details

This appendix provides the implementation details of BrainFlow, including EEG feature construction, numerical preprocessing, Healthy Velocity Field Modeling, HC-only quality control, trajectory matching, Flow-Energy computation, and prediction.

A.1 EEG Feature Construction

For each participant, we filter the retained non-overlapping 5-s EEG epochs into five frequency bands: delta (1–4 Hz), theta (4–8 Hz), alpha (8–13 Hz), beta (13–30 Hz), and low gamma (30–45 Hz). Sub-band filtering uses the IIR filtering routine in MNE-Python (Gramfort et al., 2013), followed by a Hilbert transform along the time axis. No additional tapering or boundary trimming is applied beyond the filtering routine's internal handling.

Let $z_i^{(e,b)}(t)$ denote the analytic signal for channel i, epoch e, and frequency band b. The weighted phase-lag index (wPLI) (Vinck et al., 2011) between channels i and j is computed as

$$\mathrm{wPLI}_{ij}^{(e,b)} = \frac{\left|\mathrm{mean}_t \,\mathrm{Im}\left\{z_i^{(e,b)}(t) z_j^{(e,b)}(t)^*\right\}\right|}{\mathrm{mean}_t \left|\mathrm{Im}\left\{z_i^{(e,b)}(t) z_j^{(e,b)}(t)^*\right\}\right|}. \tag{12}$$

Here, * denotes complex conjugation, and the temporal mean is taken within each epoch. Before evaluating the ratio, each channel's analytic signal is normalized by its root-mean-square magnitude across retained epochs and samples. This positive rescaling leaves the ratio unchanged apart from numerical thresholding, and numerically negligible denominators are assigned zero.

We obtain the participant-level connectivity matrix by taking the elementwise median across retained epochs:

$$W_{ij}^{(b)} = \mathrm{median}_e \left(\mathrm{wPLI}_{ij}^{(e,b)} \right). \tag{13}$$

Each 19×19 matrix is symmetric with a zero diagonal. We partition the common montage into five scalp regions, as listed in Table 3.

Table 3: Channel groups used to construct regional wPLI features.

Region	Channels	Count
Frontal	Fp1, Fp2, F7, F3, Fz, F4, F8	7
Central	C3, Cz, C4	3
Temporal	T7, T8	2
Parietal	P7, P3, Pz, P4, P8	5
Occipital	O1, O2	2
Total		19

Channel-wise node strength. For each frequency band, node strength summarizes the mean connectivity of each channel to the other 18 channels:

$$s_i^{(b)} = \frac{1}{18} \sum_{\substack{j \in \mathcal{V} \\ j \neq i}} W_{ij}^{(b)}, \tag{14}$$

where $\mathcal{V}$ contains the 19 channels in Table 3. This produces 19 node-strength features per band and 95 across the five bands.

Regional connectivity. For two distinct regions A and B, the regional descriptor is the mean of all $|A||B|$ cross-region edges. For a single region A, the within-region descriptor is the mean of the $\binom{|A|}{2}$ unique off-diagonal edges. Table 4 lists the resulting 15 regional descriptors, yielding 75 features across the five bands. Regional averaging is applied after participant-level median aggregation.

Table 4: Regional wPLI descriptors and their constituent channel-pair counts. The 15 groups partition all 171 unique channel pairs.

Regional descriptor	Channel pairs
Frontal–frontal	21
Frontal–central	21
Frontal–temporal	14
Frontal–parietal	35
Frontal–occipital	14
Central–central	3
Central–temporal	6
Central–parietal	15
Central–occipital	6
Temporal–temporal	1
Temporal–parietal	10
Temporal–occipital	4
Parietal–parietal	10
Parietal–occipital	10
Occipital–occipital	1
Total	171

Feature groups. Within each band, regional connectivity and node strength form separate groups, resulting in ten groups and $5 \times (15 + 19) = 170$ candidate features. BrainFlow fits an independent healthy velocity field to each group. Each participant contributes one subject-level feature vector per group rather than treating individual epochs as independent training samples. The 171 channel-pair values in each band are intermediate quantities and are not separately supplied to the velocity-field models.

A.2 Training-Partition Numerical Preprocessing

All numerical preprocessing is fitted within the designated training partitions and then applied unchanged to test participants. Before Healthy Velocity Field Modeling, candidate features are screened using only normative-training HCs. A feature is excluded if its missing rate exceeds 20%, its variance is non-finite or below 10^{-12}, or more than 20% of observations have an absolute training-set z-score above 12. This initial numerical screening is independent of HCQC-A.

Normative-training HCs receive source weights proportional to $n_s^{-1/2}$, where n_s denotes the number of normative-training HCs in dataset s. These source weights are combined with robust-core weights derived from five-fold cross-fitted, median/MAD-standardized feature deviations. The robust-core weights decrease from 1 to 0.20 between the 80th and 95th percentiles of the combined within-dataset and global deviation ranks. We then normalize the weights within each dataset to preserve its source-balanced total weight. This procedure downweights atypical observations without excluding participants.

Feature means and standard deviations are estimated from normative-training HCs using the combined weights. Missing values are replaced by the fitted training means before standardization. The same transformation is subsequently applied to normative-calibration, classifier-training, and test participants.

A.3 Healthy Velocity Field Modeling

Parameter fitting uses the 472 normative-training HCs with the sampling weights described in Appendix A.2. These participants are disjoint from those assigned to normative calibration, classifier training, and testing. Checkpoint selection uses the non-test HC calibration pool. Neither MDD participants nor test participants contribute to healthy velocity-field fitting or checkpoint selection.

A.4 Healthy Reference-Bank Construction

Each feature group has an independent healthy reference bank generated by its learned velocity field. For a group of dimension d_g, Gaussian starts $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_{d_g})$ are restricted to the typical-radius region

$$\left| \frac{\|\mathbf{z}\|_2^2 - d_g}{\sqrt{2d_g}} \right| \leq 2. \tag{15}$$

For each group, we construct a bank of 4,096 accepted starts. Their healthy endpoints are obtained by integrating Eq. 2 with 50 explicit Euler steps over $t \in [0, 1]$, using $\Delta t = 0.02$. The resulting bank remains fixed and is reused for all participants within the same model run.

A.5 HC-Only Normative Quality Control

HCQC-A screens candidate features according to how well the learned healthy reference represents eligible HC observations. Healthy velocity fields are fitted on normative-training HCs, whereas feature quality is evaluated using HCs from the normative-calibration and classifier-training partitions. All MDD and test participants are excluded from feature screening.

Generated-reference normal scores. HCQC-A uses feature-wise normal scores under the generated healthy reference distribution, which are distinct from the transport residuals in Section 3.3. For each group, $N_{\mathrm{ref}} = 4096$ reference samples are generated in standardized feature space from independent standard Gaussian starts using 50 Euler steps.

Let $g_{\ell j}$ denote coordinate j of generated sample ℓ. Its smoothed empirical mid-CDF is

$$F_j(a) = \frac{\sum_{\ell=1}^{N_{\text{ref}}} \mathbb{I}(g_{\ell j} < a) + \frac{1}{2}\sum_{\ell=1}^{N_{\text{ref}}} \mathbb{I}(g_{\ell j} = a) + \frac{1}{2}}{N_{\text{ref}} + 1}. \tag{16}$$

For a standardized HC feature value x_{ij}, the corresponding normal score is

$$r_{ij}^{\text{HC}} = \Phi^{-1}\left(\text{clip}\left[F_j(x_{ij}), \epsilon, 1-\epsilon\right]\right), \qquad \epsilon = 10^{-6}, \tag{17}$$

where Φ^{-1} denotes the inverse standard normal CDF. Generated-reference scores $r_{\ell j}^{\text{ref}}$ are obtained by applying the same transformation to $g_{\ell j}$. Each eligible HC or generated sample therefore contributes one score per feature.

Quality score and retention rule. Four components assess location agreement, scale agreement, generated-reference moments, and the frequency of extreme HC scores:

$$q_{\mu,j} = \exp\left(-\frac{\left|\mu(r_j^{\text{HC}}) - \mu(r_j^{\text{ref}})\right|}{0.5}\right), \tag{18}$$

$$q_{\sigma,j} = \exp\left(-\frac{\left|\log\left[\sigma(r_j^{\text{HC}})/\sigma(r_j^{\text{ref}})\right]\right|}{0.5}\right), \tag{19}$$

$$q_{g,j} = \frac{1}{2}\left[\exp\left(-\frac{|\mu(g_j)|}{0.5}\right) + \exp\left(-\frac{|\sigma(g_j) - 1|}{0.5}\right)\right], \tag{20}$$

$$q_{\text{stab},j} = 1 - \frac{1}{n_{\text{HC}}}\sum_{i=1}^{n_{\text{HC}}} \mathbb{I}\left(|r_{ij}^{\text{HC}}| > Q_{0.99}\left(|r_j^{\text{ref}}|\right)\right). \tag{21}$$

Here, μ and σ denote the empirical mean and sample standard deviation, respectively. The quantity n_{HC} is the number of eligible calibration HCs, and $Q_{0.99}$ denotes the empirical 99th percentile. The scale-agreement component is set to zero when either standard deviation is non-finite or no greater than 10^{-8}.

The final normative-quality score averages the four components:

$$Q_{\text{HC},j} = \frac{1}{4}\left(q_{\mu,j} + q_{\sigma,j} + q_{g,j} + q_{\text{stab},j}\right). \tag{22}$$

Features with $Q_{\text{HC},j} \geq 0.80$ are assigned to class A and retained; all remaining features are excluded. Disease-separation statistics and group-level fidelity diagnostics do not enter the retention rule.

The feature mask is estimated independently for each run and then applied unchanged to its test participants. HCQC-A determines the coordinates used for reference matching and detail-level Flow-Energy computation, while the fitted velocity fields retain their original group dimensions. If a group contains no retained coordinates, it is omitted from group-level aggregation without relaxing the screening threshold.

A.6 Trajectory Matching and Relative Deviation

For each participant, we select the eight nearest healthy endpoints according to the distance in Eq. 3. Matching uses HCQC-A-retained coordinates, whereas reference generation and velocity-field evaluation operate in the full feature space of each group.

The paired linear paths in Eq. 4 are evaluated at nine uniformly spaced transport times,

$$t_m = 0.05 + \frac{m-1}{8}(0.95 - 0.05), \qquad m = 1, \ldots, 9. \tag{23}$$

All probe times receive equal weight in the subsequent Flow-Energy calculation.

A.7 HC Calibration and Flow-Energy Construction

Residual calibration uses HCs from both the normative-calibration and classifier-training partitions. Each represented dataset receives equal total weight, which is distributed equally among its HCs and their matched paths. MDD and test participants do not contribute to calibration.

At each feature coordinate and probe time, relative residuals are centered by the weighted HC median and scaled by 1.4826 times the weighted MAD. Non-finite scales or scales below 10^{-4} are replaced by the corresponding weighted standard deviation and lower-bounded by 10^{-4}. These statistics are applied through Eq. 7.

For feature j and matched path k, the path-level energy is

$$A_j^{(k)} = \frac{1}{9} \sum_{m=1}^{9} \left(\widetilde{\epsilon}_{\mathrm{rel},j}^{(k)}(t_m) \right)^2 . \tag{24}$$

After ordering the eight matched paths by increasing endpoint distance, the distance-ordered Q25 aggregation is

$$E_j = 0.25 A_j^{(2)} + 0.75 A_j^{(3)}. \tag{25}$$

This corresponds to linear interpolation at rank $1 + 0.25(8 - 1) = 2.75$. All retained coordinates within a group share the same endpoint-distance ordering.

Group-level Flow-Energy. Group-level scores summarize dependencies among retained detail-level energies. For each nonempty group g, let $\mathbf{e}_g(\mathbf{x}) = [E_j(\mathbf{x})]_{j \in \mathcal{J}_g}$ denote its energy vector. The group mean $\boldsymbol{\mu}_g$ and a diagonally shrunk, ridge-regularized covariance $\widetilde{\boldsymbol{\Sigma}}_g$ are estimated from HC calibration vectors. The squared Mahalanobis distance is

$$M_g(\mathbf{x}) = \left(\mathbf{e}_g(\mathbf{x}) - \boldsymbol{\mu}_g \right)^\top \widetilde{\boldsymbol{\Sigma}}_g^{-1} \left(\mathbf{e}_g(\mathbf{x}) - \boldsymbol{\mu}_g \right). \tag{26}$$

Using the smoothed empirical HC mid-CDF of these distances, the group score is

$$E_{\mathrm{group},g}(\mathbf{x}) = -\log \left[\mathrm{clip} \left(1 - \widehat{F}_{g,\mathrm{HC}}^{\mathrm{mid}} \left(M_g(\mathbf{x}) \right) + \varepsilon, \varepsilon, 1 \right) \right], \qquad \varepsilon = 10^{-4}. \tag{27}$$

HC observations receive equal weight when fitting the group-level scores.

Global Flow-Energy. For global aggregation, group scores are mapped to normal scores using weighted HC empirical mid-CDFs. Subject weights are proportional to $n_{\mathcal{C},s}^{-1/2}$, where $n_{\mathcal{C},s}$ denotes the number of calibration HCs in dataset s. Under the fixed dataset-role split, the corresponding dataset-specific HC median vector, estimated from the calibration HCs, is subtracted. A weighted shrinkage Mahalanobis distance followed by the same upper-tail transformation then produces E_{global}.

The final representation concatenates retained detail energies, one score for each nonempty group, and the global score, as specified in Eq. 9. All parameters used to construct the joint scores are estimated from non-test HC observations.

A.8 Classifier Preprocessing and Prediction

Each component of $\mathbf{E}_{\mathrm{all}}$ is transformed by $\log(1 + E)$ and standardized using classifier-training statistics:

$$z_{ij} = \frac{\log(1 + E_{ij}) - \mu_j^{\mathrm{train}}}{\max \left\{ 1.4826 \ \mathrm{MAD}_j^{\mathrm{train}}, 10^{-6} \right\}}. \tag{28}$$

Here, μ_j^{train} and $\mathrm{MAD}_j^{\mathrm{train}}$ denote the median and MAD of the log-transformed classifier-training values. Missing transformed values are replaced by the corresponding training median.

Source-HC site-drift attenuation is applied after standardization. Let m_{sj} denote the median standardized value of feature j among classifier-training HCs from dataset s. The final classifier input is

$$h_{ij} = \frac{z_{ij}}{\sqrt{1 + \delta_j}}, \qquad \delta_j = \mathrm{SD}_{\mathrm{pop},s}(m_{sj}). \tag{29}$$

When fewer than two datasets contribute classifier-training HCs, δ_j is set to zero. All classifier preprocessing statistics are estimated from the designated training participants and then applied unchanged to test participants.

BrainFlow uses L_2-regularized logistic regression with $C = 0.1$ and a fixed MDD probability threshold of 0.5. Classifier-training observations are weighted inversely by dataset–class frequency.

B Additional Experimental Details

This appendix describes dataset selection, EEG preprocessing, the fixed dataset-role evaluation protocol, baseline configurations, and software used in the experiments.

B.1 Dataset Selection and Channel Harmonization

We use resting-state EEG from Dortmund (Wascher et al., 2025), LEMON (Babayan et al., 2019), MODMA_128 (Cai et al., 2022), Mumtaz (Mumtaz, 2016), and TDBRAIN (van Dijk et al., 2022). We select session-1 pre-task eyes-closed recordings from Dortmund, annotation-defined eyes-closed intervals from LEMON, resting-state recordings from MODMA_128, eyes-closed recordings from Mumtaz, and session-1 eyes-closed recordings from TDBRAIN. For TDBRAIN, we retain only records labeled HC, HEALTHY, or MDD and exclude other or combined diagnostic labels.

We harmonize channel names, including the mapping of T3/T4/T5/T6 to T7/T8/P7/P8, and assign a standard 10–20 montage. All retained recordings contain the same 19 channels listed in Table 3. For Dortmund, LEMON, MODMA_128, and TDBRAIN, cleaning and average rereferencing operate on the available native EEG channels before the common montage is selected. Mumtaz recordings are first restricted to the 19 common channels and are then cleaned and average-referenced. No additional rereferencing is applied after final channel selection.

B.2 Signal Cleaning and Quality Control

Continuous EEG is resampled to 250 Hz, notch-filtered at 50 Hz, and band-pass-filtered at 1–100 Hz. Bad channels are identified using PyPREP (Bigdely-Shamlo et al., 2015). For Dortmund, LEMON, MODMA_128, and TDBRAIN, a recording is excluded when the number of persistently noisy channels reaches $\lceil 0.08N \rceil$ or the total number of flagged channels reaches $\lceil 0.20N \rceil$, where N denotes the number of usable native EEG channels. Otherwise, flagged channels are interpolated. A subsequent check on the common 19-channel montage excludes recordings with at least three persistently noisy channels or five flagged channels, without further interpolation. For Mumtaz, a single 19-channel check excludes recordings with at least five channels in either category and interpolates flagged channels in the remaining recordings.

After average rereferencing, extended Infomax ICA is applied, with the number of components determined from signal rank and explained variance. Components classified by ICLabel (Pion-Tonachini et al., 2019) as eye-blink, muscle, heartbeat, channel-noise, or line-noise artifacts are removed at a probability threshold of 0.70. ICA corrections that fail amplitude or rank checks revert to the pre-ICA signal. The resulting EEG is then filtered at 1–45 Hz.

EEG is segmented into non-overlapping 5-s epochs, and epochs exceeding recording-specific peak-to-peak thresholds estimated by autoreject (Jas et al., 2017) are removed. Recordings with less than 60 s of retained EEG are excluded. A final record-level screen checks channel coverage, finite samples and spectra, flat or clipped signals, and joint amplitude–power outliers. Robust amplitude and power statistics are estimated across preprocessed records within each dataset. Joint outliers are excluded only when accompanied by additional evidence from channel quality, epoch retention, or ICA rank; warnings that do not meet these criteria are retained.

B.3 Participant and Epoch Retention

Among the 1,277 locally indexed recordings, 944 complete preprocessing and 942 satisfy the final inclusion criteria, yielding 629 HC and 313 MDD participants. Table 5 summarizes retention by dataset.

Table 5: Retention of locally indexed recordings. Each final record corresponds to one participant.

Dataset	Indexed	Preprocessed	Final
Dortmund	608	384	383
LEMON	215	177	176
MODMA_128	40	33	33
Mumtaz	58	26	26
TDBRAIN	356	324	324
Total	1 277	944	942

Within the final cohort, epoch-level quality control retains 35,271 of 42,235 complete, non-overlapping 5-s epochs, corresponding to 176,355 s of EEG for feature extraction (Table 6).

Table 6: Epoch retention within the same 942 final participants. Cells report complete 5-s epoch counts, with total duration in seconds in parentheses.

Dataset	Before epoch rejection	Retained epochs
Dortmund	14 196 (70 980)	13 901 (69 505)
LEMON	16 764 (83 820)	10 379 (51 895)
MODMA_128	2 001 (10 005)	1 816 (9 080)
Mumtaz	1 572 (7 860)	1 516 (7 580)
TDBRAIN	7 702 (38 510)	7 659 (38 295)
Total	42 235 (211 175)	35 271 (176 355)

B.4 Evaluation Protocol

The main experiments use a fixed dataset-role split with split seed 42 and five model seeds: 42, 52, 62, 72, and 82. Subject assignments remain unchanged across runs, and MDD is treated as the positive class. ACC, recall, precision, and F1 are computed separately for MODMA_128, Mumtaz, and TDBRAIN and are then averaged equally across datasets. We report the mean and population standard deviation of these dataset-macro scores across the five model seeds.

All partitions are defined at the subject level. Participants used for Healthy Velocity Field Modeling are disjoint from those used for normative calibration, classifier training, and testing. Test participants do not contribute to parameter fitting, feature screening, residual calibration, model selection, or classifier training.

B.5 Baseline and Ablation Configurations

Comparison protocol. All competing methods are implemented and evaluated within the same experimental pipeline, with identical subject-level partitions, test participants, and evaluation metrics. All methods are provided with the same non-test training subjects. In particular, the 472 HC participants used by BrainFlow to learn the healthy velocity fields are also available to all competing methods during training. Thus, BrainFlow does not benefit from additional HC data relative to the baselines. The distinction lies only in how the training data are used: BrainFlow employs these HC subjects to establish the healthy generative reference, while each competing method uses them according to its original learning objective. Test participants are excluded from model fitting, model selection, feature screening, calibration, and classifier training. Feature-based discriminative baselines use the same 170 subject-level wPLI descriptors.

Progressive comparison. The progressive comparison in Figure 3 evaluates Base, Base+HVFM, and Base+HVFM+VTDI (Ours) under the same fixed dataset-role split. All three variants use L_2-regularized logistic regression for classification.

Base directly classifies the 170 wPLI descriptors without healthy-reference modeling or HCQC-A screening. Base+HVFM uses the learned healthy velocity fields to generate reference distributions and compute signed feature-wise tail-surprise scores. Its inputs contain all candidate coordinates, group root-mean-square summaries, and a global summary. Base and Base+HVFM use training-set mean/standard-deviation scaling and logistic regression with $C = 0.1$. Base+HVFM+VTDI (Ours) uses the relative-deviation representation and classifier preprocessing described in Appendix A.8, with the same $C = 0.1$.

Component replacement comparisons. The flow-based replacement variants in Table 2 retain the trained healthy velocity fields and use the same BrainFlow classifier preprocessing and settings. These comparisons isolate alternative residual definitions, post-processing representations, and reference-trajectory strategies within the same healthy normative framework.

Absolute residual. This variant replaces the relative residual in Eq. 6 with the subject-path residual $\mathbf{r}^{\mathrm{S},(k)}(t)$ from Eq. 5. It therefore measures residual magnitude without subtracting the matched healthy-reference residual. Healthy-reference matching, time-wise HC calibration, distance-ordered aggregation, and downstream classifier settings remain unchanged.

Generated-HC CDF. This variant retains the trained healthy velocity fields but replaces trajectory-relative deviation processing with a generated-distribution-based feature representation. For each retained coordinate, the replacement feature is the two-sided tail-surprise score

$$E_j^{\mathrm{CDF}}(\mathbf{x}) = -\log\left[\min\left\{2\min\left(F_j(x_j), 1 - F_j(x_j)\right) + \epsilon, 1\right\}\right], \qquad \epsilon = 10^{-6}. \tag{30}$$

Here, F_j denotes the smoothed generated-reference CDF defined in Appendix A.5. The replacement representation combines retained feature scores, their root-mean-square summaries within groups, and the mean group score as a global summary. The HCQC-A feature mask, classifier preprocessing, and classifier settings remain the same as in BrainFlow.

Matched endpoint distance. This variant retains the same healthy reference bank and eight nearest-reference matches, but replaces relative velocity residuals with feature-wise differences between subject and matched-reference endpoints. These signed differences are standardized using HC median/MAD statistics, squared, and aggregated using the same distance-ordered Q25 rule. No velocity-field evaluation is performed during deviation computation; HCQC-A screening and the downstream classifier pipeline remain unchanged.

Random-8 Trajectories. This variant replaces subject-specific nearest-reference selection with eight healthy references sampled randomly with replacement from the same reference bank. The relative-residual definition, HC calibration, and downstream classifier settings remain unchanged.

Diffusion-based healthy reference. The Diffusion variant replaces flow matching with diffusion-based healthy-reference modeling while retaining the same normative-training HCs, feature standardization, and HCQC-A mask. Its noise predictor contains two 128-unit hidden layers and a 32-dimensional time embedding. The optimizer, training epochs, batch size, learning rate, and weight decay match the healthy velocity-field settings. With a 100-step cosine schedule and 25 deterministic DDIM sampling steps, the model generates 4,096 healthy references per group and selects the eight nearest references for each subject. Relative deviations compare subject and reference noise-prediction residuals at ten noise levels using four shared noise draws, followed by HC calibration and distance-ordered Q25 aggregation. Classifier preprocessing and settings remain the same as in BrainFlow.

B.6 SOFTWARE

EEG preprocessing and feature extraction use Python 3.9.23, MNE-Python 1.8.0, PyPREP 0.5.0, autoreject 0.4.3, mne-icalabel 0.7.0, NumPy 1.26.4, and SciPy 1.13.1.

C Additional Experiments and Analyses

This appendix reports supplementary analyses that examine feature retention, parameter sensitivity, and physiological attribution.

C.1 HCQC Feature Retention Analysis

Under the fixed dataset-role split and model seed 42, HCQC-A retains 153 of the 170 candidate features. Table 7 lists the excluded coordinates, allowing the retained feature set to be reconstructed from the definitions in Appendix A.5.

Table 7: Feature retention under the fixed dataset-role split with model seed 42. All candidates not listed in the final column are retained.

Band	Feature type	Before	After	Excluded features
Delta	Regional pair	15	15	None
Delta	Node strength	19	18	O1
Theta	Regional pair	15	15	None
Theta	Node strength	19	19	None
Alpha	Regional pair	15	11	Central–parietal; central–occipital; central–temporal; temporal–occipital
Alpha	Node strength	19	17	T7; C3
Beta	Regional pair	15	13	Frontal–frontal; central–temporal
Beta	Node strength	19	14	Fz; F3; C3; F8; T8
Low gamma	Regional pair	15	15	None
Low gamma	Node strength	19	16	F4; Fp2; P8
Total		170	153	17 features excluded

Across the five model seeds, HCQC-A retains 146–159 features, while all ten feature groups remain nonempty in every run (Table 8). Subject assignments remain fixed, whereas the learned healthy reference and the resulting feature mask vary across model seeds.

Table 8: Retained feature counts across five model seeds under the fixed dataset-role split. Each run starts from 170 candidates in ten groups.

Model seed	42	52	62	72	82
Retained features	153	152	146	159	156
Nonempty groups	10	10	10	10	10

C.2 Sensitivity to Matched-Trajectory Parameters

We examine sensitivity to two parameters in matched-trajectory inference: the number of matched healthy trajectories K and the distance-ordered aggregation quantile. Both analyses use the same fixed dataset-role split and five model seeds as the main experiments. The remaining hyperparameters follow the settings used in Flow Mismatching (Chen et al., 2026). We vary one parameter at a time while keeping all other settings unchanged.

(1) Number of matched trajectories. We evaluate $K \in \{1, 4, 8, 12\}$. For $K = 1$, the nearest reference trajectory is used directly; for larger K, distance-ordered Q25 aggregation is retained.

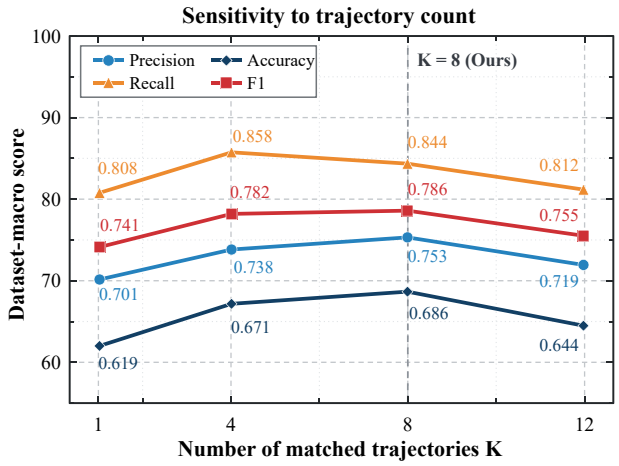

Figure 7: Sensitivity to the number of matched trajectories K under the fixed dataset-role split. Curves show dataset-macro mean ACC, recall, precision, and F1 over five model seeds. The vertical axis uses percentages, while point labels report the corresponding values on a 0–1 scale. Standard deviations are not shown. The dashed vertical line marks the default setting, $K = 8$.

As shown in Figure 7, $K = 8$ achieves the highest mean ACC, precision, and F1 among the evaluated settings, whereas recall peaks at $K = 4$. Increasing K from 1 to 8 improves all four metrics, but further increasing it to 12 reduces performance across all metrics. These results support $K = 8$ as the default setting within the evaluated range, suggesting a balance between healthy-reference diversity and subject-specific matching.

(2) Distance-ordered aggregation quantile. With $K = 8$ fixed, we compare Q25, Q50, and Q75, corresponding to aggregation quantiles of 0.25, 0.50, and 0.75, respectively. Q25 is the default configuration used by BrainFlow in the main experiments.

Table 9: Sensitivity to the aggregation quantile under the fixed dataset-role split ($K = 8$). Values are mean $\pm$ standard deviation over five model seeds. Best means are bold.

	Classification metrics			
Aggregation quantile	Pre↑	Rec↑	Acc↑	F1↑
Q25 (default)	**0.753 ± 0.025**	**0.844 ± 0.036**	**0.686 ± 0.026**	**0.786 ± 0.019**
Q50	0.694 ± 0.024	0.757 ± 0.063	0.600 ± 0.039	0.717 ± 0.038
Q75	0.744 ± 0.020	0.835 ± 0.063	0.674 ± 0.025	0.776 ± 0.028

As reported in Table 9, Q25 achieves the highest mean performance across all four metrics. Relative to Q50, Q25 improves mean ACC and F1 by 0.086 and 0.069, respectively. Q75 performs better than Q50 and approaches Q25, but its mean ACC and F1 remain lower by 0.012 and 0.010, respectively. Thus, performance does not vary monotonically with the aggregation quantile. These comparisons support retaining Q25 as the default among the evaluated quantiles under the fixed dataset-role split.

C.3 Additional Physiological Attribution Analysis

This analysis complements Sections 4.5 and 4.6 with feature-level deviation statistics and an individual attribution example.

C.3.1 Feature-Level Deviation Statistics

We examined feature-level Flow-Energy in 156 MDD and 22 HC test participants under the fixed dataset-role split (model seed 42). For each retained detail feature, the deviation threshold was the

empirical 95th percentile among 135 calibration HCs (112 normative-calibration and 23 classifier-training participants):

$$T_j = Q_{0.95}(\{E_j(x_i) : x_i \in \mathcal{H}_{\text{cal}}\}), \qquad A_j(x) = \mathbf{1}\{E_j(x) > T_j\}, \tag{31}$$

where $E_j(x)$ denotes feature-level Flow-Energy and $\mathcal{H}_{\text{cal}}$ denotes the calibration HC set. Test participants were excluded from threshold estimation. These thresholds flag elevated deviations for interpretation rather than determine classifier predictions.

Figure 8 shows the six features with the highest flag prevalence among MDD test participants. Bars indicate the percentage of participants exceeding each feature-specific threshold, and right-hand annotations report median Flow-Energy for MDD and HC. Theta-band F4 node strength and frontal–parietal connectivity showed higher flag prevalence in MDD than HC (22.4% versus 4.5% and 9.1%, respectively). In contrast, frontal–frontal theta and frontal–occipital low-gamma features showed slightly higher prevalence in HC. These feature-level summaries complement the spatial and spectral analyses in the main text.

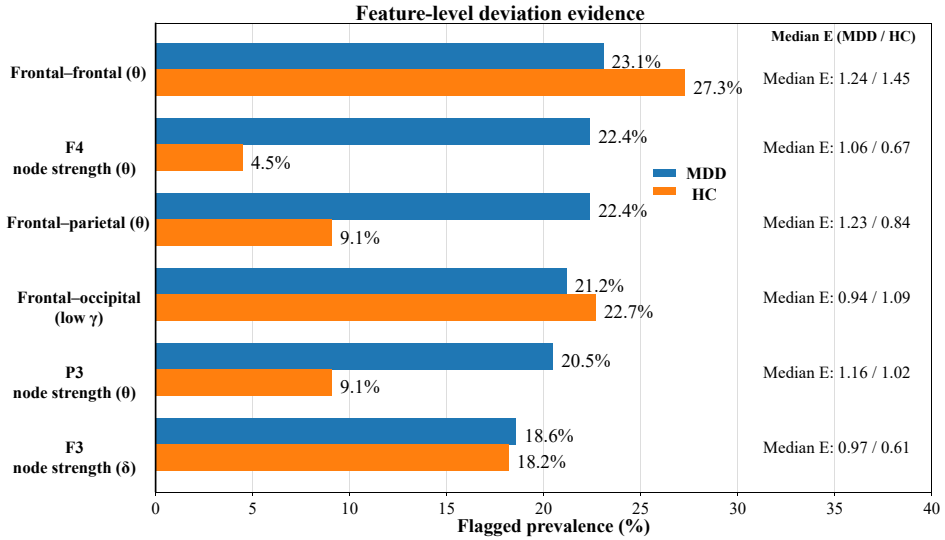

Figure 8: Feature-level deviations in 156 MDD and 22 HC test participants. The six features are ranked by MDD flag prevalence. Blue (MDD) and orange (HC) bars show percentages exceeding the feature-specific calibration-HC 95th percentile. Right-hand annotations report median Flow-Energy (MDD / HC).

C.3.2 Individual-Level Attribution Example

We illustrate the attribution procedure using one TDBRAIN test participant with a recorded MDD diagnosis under the fixed dataset-role split. This participant does not contribute to model training or calibration. BrainFlow assigns an MDD probability of 0.863, corresponding to an MDD prediction at the decision threshold of 0.5. The classifier uses 164 inputs: 153 retained detail features, 10 group-level scores, and one global score.

To distinguish deviation magnitude from prediction direction, we decompose the classifier logit in Eq. 11 into signed feature contributions:

$$p_{\text{MDD}}(x) = \sigma\left(b + \sum_{j=1}^{164} c_j(x)\right), \qquad c_j(x) = w_j h_j(x), \tag{32}$$

where $h_j(x)$ denotes the classifier input after the preprocessing described in Appendix A.8, w_j denotes its learned coefficient, b is the intercept, and σ is the logistic function. Positive contributions increase the MDD logit, whereas negative contributions decrease it. Thus, exceeding an HC reference threshold and supporting an MDD prediction represent two distinct properties.

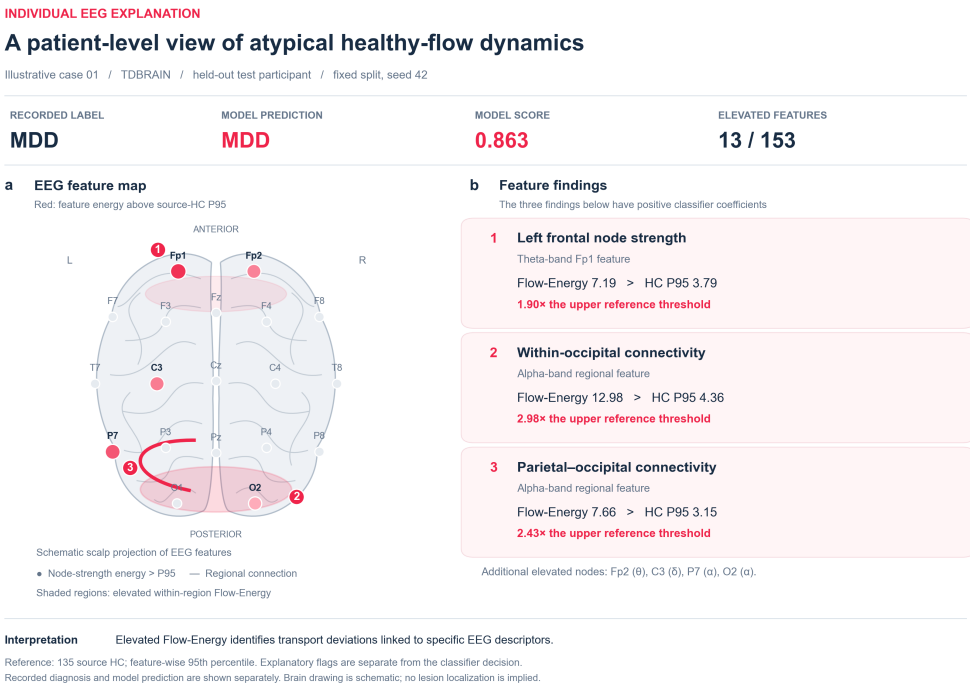

Figure 9: Individual-level attribution for a TDBRAIN test participant with a recorded MDD diagnosis under the fixed dataset-role split. BrainFlow assigns an MDD probability of 0.863, and 13 of 153 detail features exceed their HC reference thresholds. The schematic highlights selected elevated features and their Flow-Energy relative to the corresponding thresholds. Spatial annotations represent EEG descriptors rather than anatomical lesions or causal effects.

Table 10: Elevated detail features in the individual example, ordered by E_j/T_j. E_j denotes Flow-Energy, T_j the HC 95th-percentile threshold, and c_j the signed contribution to the MDD logit.

Feature	E_j	T_j	E_j/T_j	c_j
Within-occipital alpha	12.98	4.36	2.98	+0.463
Parietal–occipital alpha	7.66	3.15	2.43	+0.491
Within-frontal alpha	9.11	4.69	1.94	+0.161
Fp1 theta node strength	7.19	3.79	1.90	+0.176
P7 alpha node strength	7.62	4.61	1.65	+0.029
Central–occipital low-gamma	5.54	3.60	1.54	−0.982
C3 delta node strength	4.22	3.01	1.41	−0.189
Frontal–central delta	5.33	3.81	1.40	+0.519
Central–occipital theta	5.32	3.94	1.35	+0.481
Fp2 theta node strength	4.92	3.65	1.35	+0.154
Within-occipital theta	6.57	5.25	1.25	+0.322
Parietal–occipital low-gamma	3.21	2.78	1.16	+0.137
O2 alpha node strength	5.18	4.85	1.07	+0.063

For this participant, 13 of the 153 detail features exceed their HC reference thresholds. Figure 9 visualizes selected elevated features, and Table 10 reports all 13 features ordered by their threshold ratios E_j/T_j. The largest ratios occur for within-occipital alpha connectivity (2.98) and parietal–occipital alpha connectivity (2.43), both of which contribute positively to the MDD logit. In contrast, central-occipital low-gamma connectivity exceeds its HC threshold but contributes negatively to the MDD logit, illustrating why deviation magnitude should be interpreted together with signed

prediction contributions. The final prediction uses all 164 inputs and the intercept rather than only the flagged features.

Together, the cohort-level statistics and the individual example connect healthy-reference deviations to model predictions at complementary levels. They provide traceable evidence for BrainFlow predictions without treating individual EEG descriptors as independently validated diagnostic biomarkers.