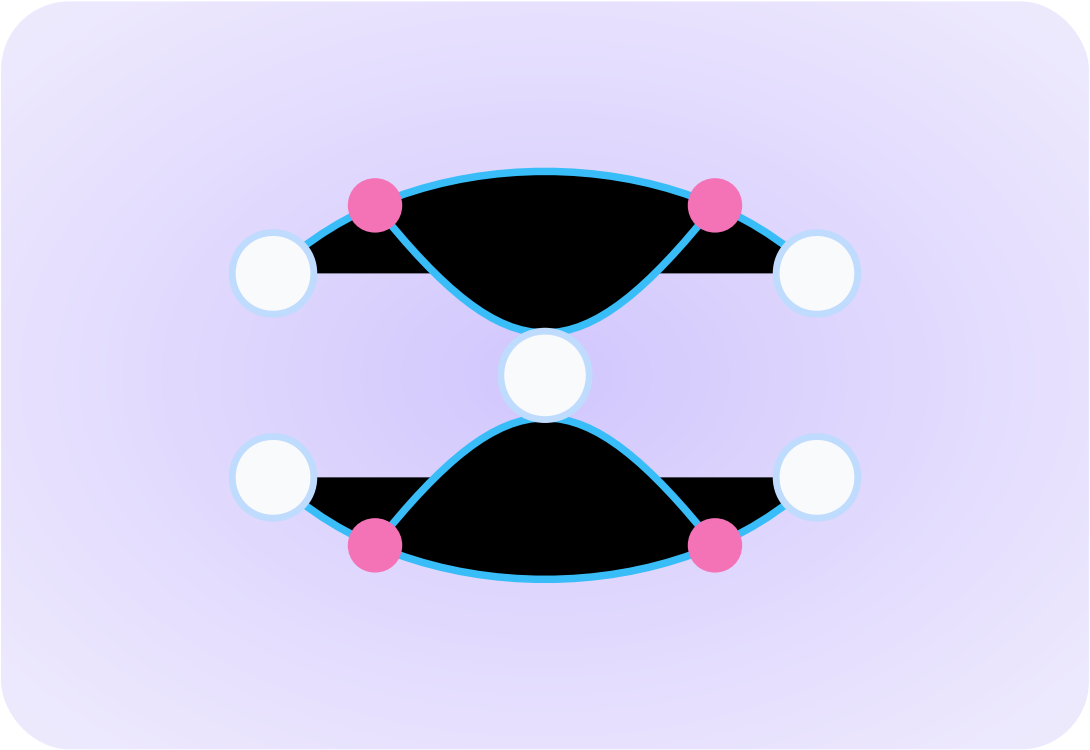
Brain Harmony: A Multimodal Foundation Model Unifying Morphology and Function into 1D Tokens



<h3>Brain Harmony: A Multimodal Foundation Model Unifying Morphology and Function into 1D Tokens · Concept Sketch</h3>
Neural dynamics lens highlighting connectivity vs. representation trade-offs.

Brain Harmony: A Multimodal Foundation Model Unifying Morphology and Function into 1D Tokens

Authors: Zijian Dong, Ruilin Li, Joanna Su Xian Chong, Niousha Dehestani, Yinghui Teng, Yi Lin, Zhizhou Li, Yichi Zhang, Yapei Xie, Leon Qi Rong Ooi, B.T. Thomas Yeo, Juan Helen Zhou
Year: 2025
Venue: NeurIPS 2025

---

1. Classification

- **Domain Category:**
  - Brain FM + Multimodal / Integration
    Brain Harmony (BrainHarmonix) is a brain foundation model that jointly models structural MRI and functional MRI, explicitly targeting multimodal integration of brain morphology and dynamics.
- **FM Usage Type:**
  - Core FM development + Multimodal FM or cross-modal integration
    The paper introduces a new multimodal brain FM architecture and training scheme, rather than applying an existing FM.
- **Key Modalities:**
  - T1-weighted structural MRI (cortical morphology).
  - Resting-state fMRI time series (functional dynamics, heterogeneous TRs).
  - Derived: geometric harmonics on cortical surface meshes, functional gradients, ROI-level (Schaefer-400) time series.

---

2. Executive Summary

Brain Harmony (BrainHarmonix) is a multimodal brain foundation model that learns a compact sequence of 1D “brain hub” tokens summarizing both structural MRI and functional MRI of the human brain. The model first trains separate encoders for T1 anatomy and fMRI dynamics, then fuses their latent representations through shared hub tokens that are optimized to reconstruct both modalities, embodying the neuroscience idea that structure constrains function. To handle real-world fMRI data collected with different temporal resolutions (TRs), the authors introduce Temporal Adaptive Patch Embedding (TAPE), which adapts patch sizes and embedding filters so that each token always corresponds to a consistent time duration, and uses multi-TR downsampling as an effective data augmentation. They also incorporate geometric harmonics derived from cortical surfaces as position embeddings for fMRI, aligning functional representations with cortical geometry. Pretrained on large UK Biobank and ABCD datasets, BrainHarmonix is evaluated on six benchmarks (ASD, ADHD, PD, MCI, cognition) plus an Asian clinical cohort and consistently outperforms strong structure-only and function-only baselines, including prior brain

FMs like BrainLM, Brain-JEPA, BrainMVP, and BrainMass. For a new grad student, this work is a concrete example of how to design a multimodal, large-scale brain FM that respects neuroscience principles, deals with heterogeneous acquisition, and yields versatile representations usable for many downstream tasks.

---

3. Problem Setup and Motivation

Scientific / practical problem:

- Learn a **unified representation** of brain structure and function that supports diverse downstream tasks (disease classification, cognition prediction) across datasets and sites.
- Existing brain FMs either model only T1 structure or only fMRI dynamics, missing complementary information and known structure–function relationships.
- fMRI FMs usually assume a single TR, limiting their ability to leverage multiple datasets and protocols and ignoring important temporal details when forced to downsample.

Why this is hard:

- **High-dimensional, heterogeneous inputs:** 3D T1 volumes and long fMRI time series are large and noisy, and come from multiple scanners, sites, and TRs.
- **Multimodal alignment:** Structural and functional images differ in resolution and sampling; mapping them into a shared latent space requires careful design to avoid one modality dominating.
- **Physics and biology constraints:** Neuroscience suggests functional waves are constrained by cortical geometry, but most models do not encode this.
- **Generalization and robustness:** The model must work across age ranges (children vs adults), disorders, and acquisition protocols, while being resistant to motion and site artifacts.
- **Interpretability and clinical relevance:** Representations should connect to known networks and clinical markers, not just achieve raw predictive performance.

---

4. Data and Modalities

Pretraining data:

- **UK Biobank (UKB):**
  - 43,112 participants (ages 44–83).
  - 46,455 T1-weighted MRI scans.
  - 40,162 resting-state fMRI time series (TR = 0.735 s).
  - fMRI downsampled to additional TRs of 1.47, 2.205, 2.94 s for multi-TR augmentation.
- **ABCD:**
  - 11,221 children (8–11 years, baseline + 2-year follow-up).
  - 18,139 T1-weighted images.
  - 30,771 resting-state fMRI time series (TR = 0.8 s) downsampled to TRs 1.6 and 2.4 s.

- **Totals:**
  - T1 pretraining: 64,594 images.
  - fMRI pretraining (after augmentation): 252,961 time series.
  - Multimodal fusion (T1–fMRI pairs): 69,360 paired sessions.

Downstream benchmarks:

- **Neurodevelopmental disorders:**
  - ABIDE-I & ABIDE-II: Autism Spectrum Disorder (ASD) vs controls (multi-site, heterogeneous TRs).
  - ADHD-200: ADHD vs controls (multi-site, heterogeneous TRs).
- **Neurodegenerative disorders and cognition:**
  - PPMI: 4-way classification (controls, SWEDD, prodromal, PD).
  - ADNI: CN vs MCI classification.
  - HCP-A: Regression of executive function (Flanker task).
- **Additional cohort:**
  - MACC (Asian clinical cohort): amyloid-positive vs negative classification.

Preprocessing / representation:

- T1: skull-stripping (FreeSurfer), reorientation (FSL), registration to MNI152 (FLIRT), cropping, intensity normalization.
- fMRI: motion correction, slice-timing correction, nuisance regression (global, white matter, CSF, motion), censoring high-motion frames, band-pass filtering, mapping to standard space or surface, then parcellation to 400 ROIs (Schaefer-400).
- Final representation: T1 patch tokens; ROI-wise fMRI time series; geometric harmonics computed on a population-average cortical mesh.

---

5. Model / Foundation Model

Model type:

- Transformer-based **multimodal foundation model** with:
  - A 3D Masked Autoencoder (MAE) for T1 (BrainHarmonix-S).
  - A JEPA-style fMRI dynamics encoder (BrainHarmonix-F) with geometric harmonics and TAPE.
  - A multimodal "Harmonizer" transformer using learnable 1D hub tokens as a shared bottleneck.

New FM vs existing:

- BrainHarmonix is a **new FM** that:
  - Encodes structure–function coupling via geometric harmonics.
  - Handles arbitrary TRs with TAPE.
  - Learns unified 1D hub-token representations that can reconstruct both modalities.

Key components and innovations:

- **BrainHarmonix-S (structure encoder):**
  - Backbone: ViT-B MAE on T1 volumes (patch size 16, ~1200 tokens).
  - Objective: reconstruct masked patches, capturing rich cortical morphology.

- **BrainHarmonix-F (functional encoder):**
  - Backbone: ViT-B with JEPA-style masked prediction over fMRI tokens.
  - **Geometric harmonics positional encoding:** eigenmodes of Laplace–Beltrami operator on a population-average cortical surface, downsampled to ROIs and linearly projected to positional embeddings.
  - **TAPE:** define a canonical temporal window τ, adjust patch size (k /), resize embedding filters via pseudoinverse operations, pad shorter sequences with attention masks—allowing consistent token semantics across TRs and enabling multi-TR augmentation.
- **Multimodal fusion (Harmonizer + hub tokens):**
  - Introduce NH learnable hub tokens, shared across all T1–fMRI pairs.
  - Concatenate hub tokens with modality-specific latents and feed through a transformer; self-attention lets hub tokens aggregate cross-modal information.
  - Lightweight decoders map hub tokens back to structural and functional latent spaces; training minimizes reconstruction error for both modalities, yielding a compact, shared latent space.
  - For downstream tasks, average-pool hub tokens and pass through a simple projection head.

Training setup:

- Encoders and Harmonizer use ViT-B backbones with FlashAttention; optimization via AdamW with cosine learning-rate and weight-decay schedules.
- Pretraining runs on 8×H100 GPUs (80 GB), with fusion training (Harmonizer) taking ~10 hours for NH=128 tokens.
- Downstream: encoders are frozen; only Harmonizer and a linear head are tuned or even just the linear head for linear probing.

---

6. Multimodal / Integration Aspects (If Applicable)

Multimodal integration in BrainHarmonix:

- **Modalities integrated:** T1 morphology, fMRI dynamics (multi-TR), and geometry-derived harmonic modes.
- **Integration strategy:**
  - **Late fusion of latent representations:** T1 and fMRI encoders are trained separately and then frozen, preserving modality-specific features.
  - **Structure-informed functional encoding:** geometric harmonics anchor fMRI ROI embeddings to cortical geometry, embedding structure–function constraints directly into functional latents.
  - **Hub-token bottleneck:** learnable 1D tokens sit between modalities and are optimized to reconstruct both structural and functional latents, forming a compact joint embedding space.

Why this integration is useful:

- Encodes the neuroscience principle that **function follows structure**, helping cross-subject and cross-dataset alignment.
- Exploits complementary strengths: stable anatomical variation (atrophy, cortical thickness) from T1 plus dynamic network organization from fMRI.

- TAPE + multi-TR augmentation make it practical to fuse data from many scanners and protocols, which is essential for large-scale multimodal foundation models.

Relation to the integration baseline plan:

- The plan emphasizes **late integration** and preserving modality-specific signal; BrainHarmonix follows this by training separate unimodal encoders and fusing only at the level of latents via hub tokens, with reconstruction losses that discourage modality collapse.
- The Harmonizer's learned joint embedding is analogous to a nonlinear CCA-style latent space, but constrained by physics-informed geometry, echoing the plan's suggestion to use CCA/structured methods before heavy fusion.
- Evaluation practice (multiple datasets, consistent splits, reporting variance and significance, ablations on fusion components) aligns with the plan's focus on robustness and disciplined comparison.
- Attention analyses reveal modality-specific and cross-modal hub tokens, consistent with the plan's goal of preserving heterogeneous modality-specific features while enabling cross-modal interactions.

---

7. Experiments and Results

Tasks / benchmarks:

- ASD vs controls (ABIDE-I, ABIDE-II), ADHD vs controls (ADHD-200). +- 4-way PD-related classification (PPMI), CN vs MCI (ADNI), executive function regression (HCP-A).
    - Amyloid-positive vs negative classification in an Asian clinical cohort (MACC).

Baselines:

- Structure-only: BrainMVP variants, BrainHarmonix-S.
- Function-only: BrainNetCNN, BrainGNN, BrainNetTF, BrainMass, BrainLM, Brain-JEPA.
- Ablations: BrainHarmonix-F, variants without geometric pre-alignment or without multi-TR augmentation, and models without multimodal fusion.

Key findings (trends):

- **Multimodal BrainHarmonix** consistently matches or outperforms all baselines across neurodevelopmental and neurodegenerative tasks, often with statistically significant gains in accuracy and F1.
- **BrainHarmonix-F** (functional-only) outperforms prior fMRI FMs (BrainLM, Brain-JEPA, BrainMass) and task-specific models, showing the value of multi-TR dynamics modeling and geometry-informed position embeddings.
- **BrainHarmonix-S** is competitive with or better than BrainMVP despite not using multi-parametric MRI, thanks to larger T1 pretraining.
- Multi-TR augmentation and geometric pre-alignment both yield consistent performance improvements; ablations confirm that removing either degrades results.
- Scaling the number of hub tokens improves performance up to ~128–256 tokens, after which gains plateau; even **linear probing** of BrainHarmonix yields

state-of-the-art or competitive results.
- On the MACC cohort, BrainHarmonix achieves the highest accuracy and F1, demonstrating promising cross-population generalization.

---

8. Strengths, Limitations, and Open Questions

Strengths:

- First brain FM to **jointly model structure and function** with a principled multimodal architecture and physics-informed inductive biases.
- Addresses a major practical barrier—**heterogeneous TRs**—via TAPE and multi-TR augmentation, enabling large-scale functional pretraining.
- Learns compact, versatile hub-token representations that support strong linear probing and efficient fine-tuning across many tasks.
- Thorough empirical validation with multiple public datasets, an independent clinical cohort, ablations, scaling studies, and interpretability analyses linking tokens to known networks and ASD-related patterns.

Limitations:

- Pretraining demography is still limited (children and middle/older adults) and mostly Western datasets; infancy, early adulthood, and broader global populations are underrepresented.
- Training requires substantial compute (8×H100 GPUs), making replication and extension difficult for small labs.
- Only T1 and resting-state fMRI are modeled; other modalities (diffusion MRI, task fMRI, EEG/MEG, genetics, clinical variables) are not yet integrated.
- Multimodal fusion can be sensitive to low-quality structural data (e.g., motion artifacts in ADHD-200 T1), which can degrade performance.
- Despite some interpretability analyses, the mapping from token-level patterns to clinically actionable insights remains preliminary.

Open questions / future directions:

1. How to extend the hub-token framework to incorporate additional modalities (diffusion, task fMRI, genomic features) while preserving modality-specific signals in line with the integration baseline plan?
2. Can joint fine-tuning of unimodal encoders and Harmonizer (possibly via parameter-efficient adapters or prompts) further improve performance without prohibitive compute?
3. How does BrainHarmonix behave in low-data or few-shot clinical settings, and can light-weight adaptation strategies close the gap?
4. Can the learned structure–function tokens support more mechanistic analyses, e.g., probing causal pathways, simulating interventions, or building “digital twin” models of individual brains?
5. What safeguards and evaluation protocols are needed before using such multimodal FMs in clinical decision support, especially concerning bias, robustness, and privacy?

---

9. Context and Broader Impact

- **Within brain FMs:** BrainHarmonix extends the line of fMRI FMs (BrainLM, Brain-JEPA) and structural FMs (BrainMVP) by **explicitly fusing structure and function**, showing that multimodal foundation models can better capture brain organization and disease-relevant signals than unimodal models.
- **Analogy to general FMs:** Conceptually, BrainHarmonix is like a CLIP-style or multimodal FM for the brain: it learns a shared token-based representation over multiple modalities (anatomy + dynamics), but with strong physics-informed priors (geometric harmonics) and careful handling of temporal sampling.
- **Relevance to multimodal integration generally:** The architecture and training strategy—separate encoders, late fusion via a compact bottleneck, physics-informed embeddings, robust multi-dataset evaluation—provide a concrete template for integrating other modalities (e.g., genomics, cell imaging, clinical data) in line with the integration baseline plan.
- **For a grad student:** This paper is a valuable reference on how to (1) scale brain FMs, (2) encode domain knowledge (geometry, TR constraints) into model design, and (3) evaluate multimodal models across heterogeneous datasets and populations.

---

10. Key Takeaways (Bullet Summary)

- **Problem:** Existing brain foundation models typically handle either structural MRI or functional MRI, and struggle with heterogeneous TRs and multimodal integration, limiting their ability to capture holistic brain organization.
- **Model:** BrainHarmonix introduces a multimodal brain FM with separate T1 and fMRI encoders, geometric harmonics-based positional encoding, TAPE for arbitrary TRs, and a Harmonizer transformer with learnable 1D hub tokens that form a compact joint representation.
- **Data:** The model is pretrained on large UKB and ABCD datasets (tens of thousands of T1 and fMRI scans) and evaluated on six benchmark datasets plus an independent Asian clinical cohort.
- **Results:** BrainHarmonix consistently outperforms structure-only and function-only baselines (including BrainLM, Brain-JEPA, BrainMVP, BrainMass) across ASD, ADHD, PD, MCI, and cognition tasks, with ablations confirming the benefits of geometric pre-alignment, multi-TR augmentation, and multimodal fusion.
- **Integration:** The fusion strategy aligns with late integration principles in the integration baseline plan—preserving modality-specific encoders, using a compact bottleneck for cross-modal alignment, and enforcing reconstruction of both modalities from shared tokens.
- **Significance:** BrainHarmonix provides a strong template for future multimodal brain FMs and suggests how to design scalable, physics-informed integration architectures that are robust across datasets and acquisition protocols.
- **For future work:** Extending the model to more modalities, broader populations, and low-resource settings, and deepening interpretability and clinical pathways, are key directions for the next generation of multimodal brain foundation models.

Generated via custom pipeline · 2025-11-18