npj Systems Biology and Applications https://doi.org/10.1038/s41540-026-00699-y

Article in Press

Inferring relationships among major psychiatric disorders in a resting-state functional connectivity-informed embedding space

Received: 11 June 2025

Accepted: 23 March 2026

Cite this article as: Bai, W., Yamashita, O., Sakai, Y. *et al.* Inferring relationships among major psychiatric disorders in a resting-state functional connectivity-informed embedding space. *npj Syst Biol Appl* (2026). https://doi.org/10.1038/s41540-026-00699-y

Wenjun Bai, Okito Yamashita, Yuki Sakai & Junichiro Yoshimoto

Inferring Relationships Among Major Psychiatric Disorders in a Resting-State Functional Connectivity-Informed Embedding Space

Wenjun Bai[1*], Okito Yamashita[1,2], Yuki Sakai[3], Junichiro Yoshimoto[1,4,5]

[1]Department of Computational Brain Imaging, Advanced Telecommunication Research Institute International (ATR), Kyoto, Japan.
[2]Center for Advanced Intelligence Project, RIKEN, Tokyo, Japan.
[3]Department of Neural Computation for Decision Making, Advanced Telecommunication Research Institute International (ATR), Kyoto, Japan.
[4]Department of Biomedical Data Science, School of Medicine, Fujita Health University, Aichi, Japan.
[5]International Center for Brain Science, Fujita Health University, Aichi, Japan.

*Corresponding author(s). E-mail(s): wjbai@atr.jp;

Abstract

Major neuropsychiatric disorders such as major depressive disorder (MDD) and schizophrenia (SCZ), as well as the neurodevelopmental disorder autism spectrum disorder (ASD), are traditionally treated as distinct clinical entities. However, genome-wide association studies indicate shared genetic risks, motivating a trans-diagnostic view. Resting-state functional connectivity (rsFC) is a promising biomarker for these disorders, but its high dimensionality complicates inference of inter-disorder relationships in the native feature space. Here we develop an rsFC-based embedding-relation workflow that quantifies disorder relationships in a connectivity-informed, low-dimensional embedding space. Central to the workflow is a mutual information-based embedding framework that evaluates candidate embedding approaches and selects an optimal strategy. Using synthetic connectivity data, the framework indicates that rsFC embeddings are best

represented in a spherical space under a moderate level of supervision. Building on this insight, we applied the workflow to curated, multi-disorder rsFC datasets to derive shared embedding spaces encompassing the connectivity features of ASD, MDD, and SCZ. In these spaces, we consistently observed a robust three-way relationship: a pronounced neurobiological dissimilarity between ASD and MDD, contrasted with greater similarity between SCZ and both disorders. These findings support a dimensional, transdiagnostic perspective on neuropsychiatric disorders and offer new insights into their shared and distinct neural underpinnings.

Keywords: Resting-State functional connectivity, Psychiatric disorders, Dimensionality reduction, Mutual information embedding framework, fMRI

Introduction

Unveiling the relationships among multiple psychiatric disorders remains a long-standing challenge in both research and clinical practice. From early studies on disorder comorbidity [1, 2] to recent efforts promoting parsimonious classification in clinical domains, such as the revision of diagnostic guidelines from DSM-IV to DSM-5 [3] and the RDoC initiative [4]. Substantial efforts have been dedicated to identifying associations between disorders in pursuit of a unified, dimensional view of major neuropsychiatric conditions [5].

A series of recent genome-wide association studies [6–9] have highlighted shared genetic factors across major disorders, providing biological support for this dimensional perspective. Beyond genetic biomarkers, numerous studies have focused on neuroimaging biomarkers to uncover distinctive brain signatures of these disorders [10, 11]. Among various neuroimaging modalities, functional MRI – particularly resting-state functional connectivity (rsFC), which measures the correlation between spatially distant brain regions [12] – has become a widely used tool for characterizing multiple disorders. This has been further bolstered by large-scale, cross-institutional efforts to collect multi-disorder MRI data, including the UK Biobank [13], the RDoC initiative [4], and the Brain/MINDS Beyond project [14].

Significant progress has been made using rsFC biomarkers to identify discriminative connectome signatures for autism spectrum disorder (ASD) [15], major depressive disorder (MDD) and its subtypes [16, 17], and schizophrenia (SCZ) [18–20]. However, while rsFC has been successfully utilized for disorder prediction and stratification, quantifying neurobiological relationships, which are defined as the degree of similarity or dissimilarity among disorders, remains challenging in the native rsFC space due to its high dimensionality [21, 22]. In such high-dimensional settings, standard distance measures often become uninformative, motivating the learning of lower-dimensional representations where geometric relationships are more stable and interpretable.

Deep generative models, particularly variational autoencoders (VAE) [23], provide a principled probabilistic framework for learning such low-dimensional embeddings from high-dimensional observations. Moreover, the semi-supervised deep generative

modeling framework (SS-VAE) [24] establishes a mathematical foundation for incorporating partial-to-full supervision to structure these embedding spaces (see Table S1 for a detailed technical comparison between our and related works). Recent neuroimaging works have demonstrated the promise of VAE-based embeddings for exploring psychiatric continua and heterogeneity, for instance, the embedding space interpolation of functional connectivity patterns [25]. Nevertheless, standard VAE and SS-VAE objectives are typically optimized for data generation or label prediction; they do not enforce that embedding distances correspond to neurobiological dissimilarities. Without explicit geometry-aware constraints, an embedding may successfully separate classes yet remain poorly calibrated for interpreting the relative proximity among disorders.

To address this gap, we present a rsFC-based embedding-relation workflow designed to reveal the neurobiological relationships between disorders by extending the SS-VAE framework with explicit geometric constraints and varying amount of diagnostic information. We project the high-dimensional rsFC features into a low-dimensional embedding space, the topology of which was determined to be optimal based on its fidelity in recovering ground-truth geometries in synthetic rsFC data. The proposed workflow comprises three components: the estimation of rsFC features, a mutual information-based embedding framework (building upon [24]), and the subsequent extraction of between-disorder relationships. In contrast to prior approaches restricted to rigid embedding architectures, our framework systematically derive embedding approaches under diverse geometric constraints (e.g., spherical vs. Euclidean priors), with the flexibility to integrate subject-wise similarity and adjust the magnitude of diagnostic information.

When evaluated using synthetic rsFC data, the optimal rsFC-informed embedding approach was identified as a method with a spherical embedding space, incorporating median amount of diagnostic information. Leveraging this optimal embedding approach, we implemented the proposed embedding-relation workflow on two independent multi-disorder MRI datasets, curated from the SPRBS database as part of the Brain/MINDS Beyond project [14]. This facilitated the derivation of two rsFC-informed embedding spaces that effectively capture the connectivity features of major disorder phenotypes (ASD, MDD, and SCZ). Through analysis of these rsFC embeddings, we identified a robust three-way neurobiological relationship: while ASD and MDD exhibit minimal similarity, SCZ displays neurobiological similarity with both disorders in these connectivity-informed embedding spaces. These findings support a dimensional view of psychopathology, highlighting the continuous nature of major psychiatric phenotypes [26].

Results

Identifying the optimal embedding using synthetic rsFC data

Before analyzing real multi-disorder rsFC features, we first performed controlled experiments on synthetic rsFC data to select the optimal embedding approach within our rsFC-based embedding-relation workflow (see Methods).

In specific, we evaluated the performance of embedding approaches – `Gaussian`, `vMF`, `distance-preserving` (on either `Gaussian` or `vMF`), `least-diagnostic`, `median-diagnostic`, and `most-diagnostic` – on synthetic rsFC data to determine the optimal embedding approach. This empirical evaluation was designed to:

1. verify whether a spherical (vMF) prior are superior to conventional Gaussian prior;
2. confirm the necessity of incorporating the subject-wise similarity term for learning optimal embeddings;
3. determine the adequate amount of diagnostic information required to structure the embedding space;
4. and demonstrate qualitatively the empirical superiority of our proposed embedding approach over existing ones.

Synthetic rsFC features & ground-truth embeddings. As depicted in Fig. 1a, we employed a three-stage process to synthesize high-dimensional rsFC features with known ground-truth low-dimensional between-cluster relationships: (1) Latent Sampling: We sampled 300 2D embeddings from three distinct 2D multivariate Gaussian distributions, $\mathcal{N}_2(0.8, 1)$, $\mathcal{N}_2(0.4, 1)$, and $\mathcal{N}_2(0.2, 1)$, to form three clusters of ground-truth embeddings. (2) Manifold Learning: We learned a nonlinear projection function $\mathcal{F}$ using the computed rsFC features of 300 subjects randomly selected from the SRPBS database [27]. (3) Feature Generation: We utilized the inverse projection function $\mathcal{F}^{-1}$ to map the 300 sampled 2D embeddings from the low-dimensional ground-truth space back into the high-dimensional rsFC feature space, resulting in a synthetic rsFC feature matrix of size $[300, 9730]$.

Included embedding approaches and evaluation metrics. To determine whether the `vMF` embedding approach is superior to the `Gaussian` embedding approach, we optimized Eq. 2 by imposing Gaussian and vMF priors, respectively, on the synthetic rsFC features. Next, to examine the effect of embedding subject-wise similarity in both Gaussian and spherical embedding spaces, we optimized Eq. 4 to learn `distance-preserving` embeddings under both Gaussian and vMF conditions. Finally, to assess the impact of varying amount of diagnostic information, we optimized Eq. 5 to learn the `least-diagnostic`, `median-diagnostic`, and `most-diagnostic` embeddings. Implementation details, including network architectures and training configurations for each embedding approach, are provided in SI §2 and Table S2.

We benchmarked against a comprehensive suite of existing embedding methods, ranging from conventional linear approaches such as `linear PCA` and `local linear embedding (LLE)` [28], to graph-based methods like `Graph2V` [29], and widely used manifold learning approaches such as `t-SNE` [30] and `UMAP` [31]. Fully supervised neighborhood embedding approaches, including `NCA` [32] and the variational autoencoder-based `supervised VAE` [24], were also included. Detailed network and training configurations for these benchmarks are provided in SI §3 and Table S3.

Since our synthetic rsFC features were generated through an ill-posed low-to-high dimensional mapping ($\mathcal{F}^{-1}$), simply assessing the visual fidelity between the learned and ground-truth embeddings may not be sufficient to evaluate the quality of the derived embeddings. Therefore, we employed a multi-metric evaluation strategy. First,

we computed the mean correlation coefficient (MCC) [33] to quantify how well the embedding captures the original data structure. Second, to quantify the recovery of inter-cluster topology, we utilized the previously introduced FID metric [34] as our primary distance measure. For the ground-truth embeddings, these relationships are expressed as $\mathcal{D}_{c_1-c_3} > \mathcal{D}_{c_1-c_2} > \mathcal{D}_{c_2-c_3}$. To ensure geometric robustness, particularly for the spherical vMF embeddings, we also calculated the geodesic distance as a secondary measure of between-cluster relationships.

Based on these metrics, we evaluated inter-cluster relationships for each embedding type using a 10-fold cross-validation scheme. Across the 10 folds, we performed multiple pairwise comparisons between clusters using paired t-tests with Bonferroni correction to determine whether the ground-truth topological relationships were statistically recovered by each embedding approach (SI §3). Additional analyses, including an additional experiment on second set of synthetic rsFC data (Fig. S2), stability assessments of estimated cluster centroids (Fig. S5a) and heterogeneity evaluations (Fig. S6a), were also conducted to examine cluster overlap effects [35] and optimality of selected embedding method.

Evaluated performance of included embedding approaches. Comparison between the conventional Gaussian embeddings in the $\mathbb{R}^2$ embedding space and the proposed spherical (vMF) embeddings revealed distinct embedding performance and structures. As illustrated in Fig. 1b (upper tier), Gaussian cluster centers tend to concentrate around the origin, whereas the vMF embeddings residing on S^1 (a unit circle), resulting in qualitatively more separable clusters that more closely resemble the ground-truth embeddings. This visual observation is quantitatively supported by the mean correlation coefficient (MCC) scores (Fig. 2a), where vMF embeddings achieved a higher MCC compared to the Gaussian baseline. However, regarding the preservation of inter-cluster relationships, neither approaches successfully recovered the ground-truth between-cluster inequality (Fig. 2b-c), as indicated by the low inequality satisfaction rates (0% to 30%) in Fig. 2d. This suggests that without supervision, disorder rsFC features alone are insufficient to produce a low-dimensional space that accurately reflects both phenotype separability and relative neurobiological distances.

We subsequently attempted to guide the embedding process by leveraging estimated pairwise distances between samples. Theoretically, this should align the affinity between embeddings with the corresponding pairwise similarities in the high-dimensional rsFC space. However, contrary to expectations, determining the embeddings with pairwise distance constraints failed to improve performance. In both $\mathbb{R}^2$ and S^1 spaces (Fig. 1b, middle tier), this resulted in lower MCC scores than even the unsupervised baselines (Fig. 2a) and failed to reconstruct the correct cluster hierarchy (Fig. 2b-c; 10% satisfaction in Fig. 2d. This null effect likely stems from the unreliable computation of the high-dimensional dissimilarity matrix d^2x_{ij}, inline with the well-documented the curse of dimensionality challenge in high-dimensional distance measure [22].

In the final evaluation phase, we assessed the impact of incorporating varying degrees of diagnostic information. The introduction of diagnostic priors dramatically improved embedding quality (Fig. 2b-c). Even with the least amount of information, clusters became distinct (Fig. 1b, lower tier, left panel), yielding MCC scores ing

0.7. Most notably, the embedding with the median amount of diagnostic information emerged as the optimal configuration (Fig. 1b, lower tier, center panel). It not only achieved the highest classification accuracy (MCC $>$ 0.9; Fig. 2a) but was also the sole condition to achieve 100% satisfaction for the ground-truth inequality relation across both FID and geodesic metrics (Fig. 2d). Interestingly, while the most diagnostic information setting produced highly compact clusters (Fig. 1b, lower tier, right panel), it failed to preserve relative geometric relationships (0% satisfaction in Fig. 2d), suggesting that excessive supervision may collapse the manifold structure and compromise intrinsic inter-cluster distances.

Advantages over alternative embedding approaches. Our spherical embedding with median diagnostic information demonstrated robust embedding performance, particularly in recovering the topological structure of the data. When applied to synthetic rsFC features across 10-fold cross-validation, ten alternative embedding approaches (see Methods) struggled to consistently reproduce the ground-truth between-cluster relationships (Fig. 3), regardless of whether the embeddings were learned through unsupervised or supervised methods. While other approaches may excel in maximizing cluster compactness or class separability, our qualitative results in Fig. 3, suggest that the `median-diagnostic` approach is most effective for retrieving the specific underlying between-cluster relationships in the synthetic rsFC-informed embedding space.

Inferred relationships among ASD, MDD and SCZ in the connectivity-informed embedding space

Building on our previous evaluation using synthetic rsFC features, the optimal embedding approach was identified as the spherical (vMF) embedding incorporating median amount of diagnostic information. Leveraging this approach, we applied workflow on rsFC features that are curated from multi-disorder (ASD, MDD, and SCZ) fMRI datasets to explore their neurobiological relationships in derived connectivity-informed embedding spaces.

SRPBS Multi-disorder Database: UTO and HuShoWa Datasets. The curated resting-state functional connectivity (rsFC) features for the three neuropsychiatric disorders were obtained from the SRPBS database, a large-scale resting-state fMRI repository acquired across eight different imaging sites. This database contains ten distinct clinical phenotypes of neuropsychiatric disorders [27] and was compiled under the Brain/MINDS Beyond human brain MRI project [14]. Full details regarding the scanning protocols and curation steps are available in [27] and at https://bicr-resource.atr.jp/srpbsfc/.

At each of the eight imaging sites contributing to the SRPBS database, resting-state fMRI data were acquired in a 10-minute, eyes-open scanning session. All fMRI data underwent identical preprocessing steps, including slice-timing correction, realignment, co-registration to T1-weighted structural images, segmentation, normalization to Montreal Neurological Institute (MNI) space, and spatial smoothing with an isotropic Gaussian kernel of 6 mm full-width at half maximum. Motion parameters (6 rigid-body parameters and their first derivatives) were regressed out of

the fMRI time series. Importantly, no significant differences in mean framewise displacement (FD) were found between diagnostic groups for either dataset (ANOVA: $F_{(2,104)} = 1.21, p = 0.30$ for UTO; $F_{(2,304)} = 0.87, p = 0.42$ for HuShoWa). Global signal regression and band-pass filtering (0.01–0.1 Hz) were also applied.

- **UTO Dataset**
  From the SRPBS multi-disorder database, we curated a primary dataset, referred to here as the *UTO dataset*, which comprises resting-state fMRI scans acquired at the University of Tokyo Hospital using a GE MR750W scanner (Fig. 4). This dataset includes three clearly diagnosed disorders: ASD ($n = 10$), MDD ($n = 62$), and SCZ ($n = 35$), giving a total of 107 subjects. Since all scans from a single site, this largely mitigates the site-related variability, which is often encountered in multi-center fMRI studies.
  Following the standard preprocessing described above, rsFC features were extracted by parcellating each subject's preprocessed fMRI data using the BSA-AAL atlas [15, 36], which segments the brain into 140 regions of interest (ROIs). Pearson correlations among the time courses of these ROIs produced subject-wise functional connectivity matrices. This procedure yielded an $[107, 9730]$ feature matrix (i.e., 9730 unique ROI–ROI connections for each of the 107 subjects), which served as the input for analyses on the UTO dataset.
- **HuShoWa Dataset**
  A second dataset, referred to here as the *HuShoWa dataset*, was also derived from the SRPBS database (Fig. 4). It comprises 307 subjects acquired at two different sites: Hiroshima University (Hiroshima, Japan) with 173 MDD individuals, and Showa University (Tokyo, Japan), for participants with ASD ($n = 115$) or SCZ ($n = 19$).
  Using the same preprocessing and rsFC feature extraction methods as for the UTO dataset, we obtained an $[307, 9730]$ rsFC feature matrix. Because HuShoWa dataset was acquired using multiple scanners and protocols, we applied the empirical Bayes-based ComBat harmonization technique [37] to reduce between-site effects. This step regressed out scanner- and protocol-specific variability to some extent [38], yielding a harmonized feature matrix for subsequent analyses.

Relation quantification and validation schemes. For both the UTO and HuShoWa datasets, we utilized the FID metric to quantify the inter-disorder distances in the learned low-dimensional embedding space: $\mathcal{D}_{ASD-MDD}$, $\mathcal{D}_{ASD-SCZ}$, and $\mathcal{D}_{MDD-SCZ}$. These FID metrics reflect the topological proximity between disorders in the rsFC-informed embedding space. Consequently, we inferred the relationships between disorders based on the resulting three-way inequality observed among these pairwise distances.

Given that real-world neuropsychiatric rsFC data lack ground-truth low-dimensional representations, we implemented three distinct validation schemes to assess the reliability and robustness of the inferred relationships. These schemes included: (1) conventional 10-fold cross-validation, (2) cross-dataset validation, and (3) healthy-control validation. Implementation details are documented in SI §4 and Table S4.

1. **10-fold cross-validation**
   A standard 10-fold cross-validation procedure was applied independently to both the UTO and HuShoWa datasets to validate relationships among the three disorders. For each dataset, this scheme involved training the optimal `median-diagnostic` embedding model ten times to compute the three inter-disorder FID measures on the test set across the folds. Paired t-tests with multiple comparison corrections were then used to assess the statistical significance of pairwise differences, thereby establishing the validity of the observed three-way inequality.
2. **Cross-dataset validation**
   To evaluate the generalizability of our findings beyond a single site or acquisition protocol, we conducted a cross-dataset validation. In this framework, the UTO and HuShoWa datasets served alternately as training and testing sets. In the first round, the UTO dataset was used for training and the HuShoWa dataset for testing; in the second round, the roles were reversed. Each permutation was repeated 10 times, and the resulting FID measures were analyzed using pairwise t-tests with multiple comparison corrections to verify if the three-way inequality held across datasets.
3. **With healthy-control validation**
   To assess the stability of the inferred neurobiological relationships when the embedding space is perturbed by an additional phenotype, we introduced a healthy-control validation scheme. We integrated rsFC features from healthy participants recruited from both the UTO ($n = 170$) and HuShoWa ($n = 363$) datasets. With the inclusion of these healthy controls (HC) as a distinct class, two 10-fold cross-validation schemes were independently applied to the augmented datasets. This allowed us to determine whether the relative relationships between the disorder groups remained consistent in the presence of the healthy control population.

Inferred inter-disorder relationships in rsFC-informed embedding spaces. Irrespective of the validation schemes utilized, the low-dimensional embedding spaces derived from the UTO and HuShoWa datasets consistently produced compact and separable embeddings that aligned with clinical phenotypes. This was achievable only when leveraging spherical embedding space with median amount of diagnostic information. These disorder-separable embeddings contrast sharply with the overlapping representations produced by conventional approaches such as NCA and t-SNE (Fig. S8).

Under the 10-fold cross-validation scheme, the consistent embedding structures and high statistical power (Fig. 5; Fig. S3) revealed a robust neurobiological relations among disorders. The most striking and consistent finding across all analyses was the substantial neurobiological dissimilarity between ASD and MDD. As shown in the FID distance matrices (Fig. 5d, left), the distance between ASD and MDD ($\mathcal{D}_{ASD-MDD}$) was consistently the largest among all pairwise comparisons.Regarding the placement of Schizophrenia (SCZ) relative to these two distinct poles, SCZ consistently occupied an intermediate topological position. In the majority of validation folds (100% in FID metrics), SCZ exhibited neurobiological similarity to both disorders, typically situated closer to MDD than to ASD. This resulted in a dominant three-way inequality

expressed as:

$$\mathcal{D}_{ASD-MDD} > \mathcal{D}_{ASD-SCZ} > \mathcal{D}_{MDD-SCZ}.$$

Statistical analysis of the UTO and HuShoWa datasets confirms this hierarchy. The neurobiological dissimilarity between ASD and MDD was statistically significant, with their pairwise distance exceeding the distances of SCZ to either disorder ($\mathcal{D}_{ASD-MDD}$ vs. $\mathcal{D}_{MDD-SCZ}$: UTO $t_{10} = 3.57, p < 0.001$; HuShoWa $t_{10} = 9.84, p < 0.001$; Fig. 5d). While the primary FID metric suggests SCZ is consistently closer to MDD, we observed that using the secondary geodesic metric (Fig. S4) occasionally altered the ranking of the smaller distances (inverted the rank order between $\mathcal{D}_{ASD-SCZ}$ and $\mathcal{D}_{MDD-SCZ}$), yet the maximal dissimilarity between ASD and MDD remained constant. This confirms that while the specific affinity of SCZ may show minor metric-dependence, its role as an intermediate phenotype is topologically stable.

These inter-disorder relationships partially survived cross-dataset validation. In the second round (HuShoWa training / UTO testing), the three-way inequality was fully preserved ($\mathcal{D}_{ASD-MDD}$ vs. $\mathcal{D}_{ASD-SCZ}$: $t_{10} = 13.20, p < 0.001$). However, in the first round, the distinction between the smaller distances ($\mathcal{D}_{ASD-MDD}$ vs $\mathcal{D}_{MDD-SCZ}$) was less pronounced ($p = 0.51$; Fig. 5b), further suggesting that while the ASD-MDD separation is robust, the specific boundaries of SCZ's neurobiological similarity can be sensitive to site-specific rsFC distributions.

Finally, the healthy-control validation confirmed that these relationships are not artifacts of a closed-set classification. Even with the introduction of healthy control (HC) features, the pronounced neurobiological dissimilarity between ASD and MDD and the intermediate similarity of SCZ remained intact ($\mathcal{D}_{ASD-MDD}$ vs. $\mathcal{D}_{ASD-SCZ}$: UTO $t_{10} = 13.20, p < 0.001$; HuShoWa $t_{10} = 5.58, p < 0.001$). Notably, these relationships were consistently observed in certain alternative supervised embedding methods (e.g., UMAP, Supervised VAE; Fig. S8), reinforcing the reliability of rsFC in inferring the neurobiological relationships among these disorders.

Neuroscientific importance of the rsFC-informed embedding space

With reliable three-way relationships established, we sought to identify the functional connections that drive the geometry of the learned low-dimensional embedding space. Unlike linear models used in standard biomarker studies [16, 17], we utilizes deep neural networks with nonlinear transformations, rendering standard feature importance metrics inapplicable. To address this, we employed Layer-wise Relevance Propagation (LRP) [39] to backpropagate relevance scores through the network, thereby isolating the input rsFC features (functional connections) that most significantly influence the embedding coordinates.

Specifically, we applied the LRP-z^B rule to the trained models for both the UTO and HuShoWa datasets (SI §5) to identify the top-10 most significant nodal-level functional connections (Tables S5 and S6). The stability of these identified features was verified via permutation-based robustness validation (Fig. S9). To facilitate interpretation, we mapped these connections to their corresponding canonical networks using the Yeo-7 parcellation [40], as visualized in Fig. 6a. At the network level, we observed

a moderate degree of convergence between the datasets, with a 40% overlap in influential network-wise connections. Nodal-level analysis highlighted a consistent role for the connection between the right inferior frontal sulcus and the left temporal pole across both cohorts. Notably, the spatial distribution of these influential features revealed a consistent bias toward inter-hemispheric connectivity (comprising 6 out of the top 10 connections in both datasets), suggesting that long-range functional integration may play a pivotal role in defining the topological distinctions between these disorders. However, given the high dimensionality of the feature space (9,730 connections), we present these specific connections as illustrative of the embedding's biological basis rather than as definitive diagnostic biomarkers.

Finally, to assess the potential clinical relavance of the learned rsFC-informed embedding spaces, we investigated associations between the embedding coordinates and symptom severity scores in the HuShoWa dataset. Specifically, we examined the relationship between the composite embedding scores and the Autism Spectrum Quotient (AQ) [41] for ASD, and the Beck Depression Inventory (BDI) [42] for MDD (see SI §5 for implementation details).

Across the three cross-validation schemes, we observed directional trends linking clinical severity to the learned rsFC-informed embeddings. While the effect sizes were modest, statistically significant associations were detected in several instances, particularly for ASD phenotypes (ASD: $\overline{r} = -0.091, \rho_{cor} < 0.05$ [X-val]; $\overline{r} = -0.246, \rho_{cor} < 0.01$ [X-dataset]; $\overline{r} = -0.088, \rho_{cor} = 0.10$ [W.healthy]. MDD: $\overline{r} = 0.067, \rho_{cor} = 0.15$ [X-val]; $\overline{r} = 0.116, \rho_{cor} < 0.05$ [X-dataset]; $\overline{r} = 0.059, \rho_{cor} = 0.25$ [W.healthy]; Fig. 6b; Table S7). Although the correlations are subtle, comparative analysis indicates that our rsFC-informed embedding-relation workflow captures these clinical signals more effectively than alternative embedding approaches (Fig. 6c).

Discussion

In this study, by leveraging the proposed rsFC-based embedding-relation workflow, we successfully projected the high-dimensional rsFC features of multiple neuropsychiatric disorders onto a spherical low-dimensional embedding space with median amount of diagnostic information. Using curated rsFC data from ASD, MDD, and SCZ, we quantified the inter-disorder relationships within this rsFC-informed embedding space. Our analysis inferred a robust three-way neurobiological relationships: ASD and MDD exhibit a pronounced neurobiological dissimilarity, whereas SCZ displays an intermediate neurobiological topology, showing similarity with both ASD and MDD.

A key distinguishing feature of our workflow is its focus on producing a relation-aware embedding space. In this framework, the primary objective is to derive relationships between subject clusters (i.e., disorder phenotypes), rather than the relationship between feature clusters (e.g., brain regions) within a single subject. This contrasts with existing functional connectivity embedding workflows, such as BrainSpace [43], rest2vec [44], and hyperbolic disc embedding [45], which typically employ non-vectorized rsFC features to create "brain-aware" embeddings where each node represents a regional connection. By shifting the embedding focus from brain

regions to patient phenotypes, our directly facilitates the quantitative comparison of disorder relations.

From a dimensionality reduction perspective, the proposed mutual information-based framework in our workflow advances beyond standard approaches in two critical ways. First, unlike local neighborhood methods such as t-SNE and UMAP, which prioritize local clustering at the expense of global geometry, our framework falls into the "global" category [46]. By employing a global neighborhood that considers all samples simultaneously, we preserve pairwise distances across the entire manifold, resulting in a more faithful representation of the inter-class relationships [30]. Furthermore, while t-SNE and UMAP assume fixed data distributions (Student-t and uniform, respectively), our framework offers flexibility, supporting both Gaussian (Euclidean) and spherical (vMF) priors to match the intrinsic geometry of the data.

Second, we introduce a continuous to supervision. Unlike the traditional trichotomy of dimensionality reduction (strictly supervised, semi-supervised, or unsupervised), our workflow treats diagnostic information in a continuous manner. We advocate for the median use of diagnostic information to balance class separability with the preservation of intrinsic biological geometry. This addresses a key limitation in psychiatric modeling: diagnostic labels based solely on clinical criteria are subject to inter-rater subjectivity [47] and longitudinal variability [48]. By not fully constraining the embedding to these discrete labels (using only median amount of diagnostic information), our model allows the rsFC-informed embeddings to reveal relationships that might be obscured by rigid clinical categorization.

The observed inter-disorder relations, specifically the neurobiological dissimilarity between ASD and MDD contrasted with the similarity of SCZ to both disorders, mirrors emerging evidence from multi-modal neuroimaging and genomic studies. While ASD and MDD are frequently comorbid clinically, our results suggest they possess distinct underlying neurobiologies. This separation aligns with recent structural covariance analyses of cortical gradients, which have shown that ASD and MDD occupy distinct positions along the principal hierarchy of brain organization, whereas SCZ aligns more closely with transdiagnostic hubs [49].

Furthermore, the pivotal position of SCZ, situated between ASD and MDD in the rsFC-informed space, is consistent with multimodal fusion studies identifying SCZ as sharing distinct subnetworks with each condition: caudate-thalamus circuits shared with MDD and temporal-lingual circuits shared with ASD [50]. At the molecular level, this geometry is further supported by transcriptomic findings showing that ASD and SCZ share significant gene expression alterations (e.g., synaptic downregulation), while MDD displays a divergent profile characterized by inflammatory pathway dysregulations [7, 51]. Our rsFC-based embedding thus captures these macro-scale biological realities that may be obscured when viewing disorders solely through overlapping clinical symptoms.

The observed neurobiological dissimilarity of ASD and MDD in our embedding spaces also offers a critical perspective on clinical comorbidity. Depressive symptoms in ASD are highly prevalent but often refractory to standard MDD treatments. Our findings support the hypothesis that "depression in ASD" may be biologically distinct from primary MDD, suggesting that comorbidity in these cases represents the co-occurrence

of distinct pathologies rather than a shared etiology. This underscores the potential of our workflow to guide precision medicine: by mapping a patient's connectivity profile, clinicians might distinguish between distinct biological subtypes of depression, potentially informing more targeted therapeutic strategies beyond standard pharmacotherapy.

However, several interpretative caveats should be noted regarding these inferred inter-disorder relationships. The demonstrated neurobiological dissimilarity between ASD and MDD reflects relative differences in a coarse-grained, low-dimensional space (i.e., 2D embeddings). Interpretations beyond this embedding space should be approached with caution. Specifically, the "neurobiological dissimilarity" in the embedding space does not contradict findings of clinical comorbidity [52], nor does it imply that these disorders share few biological features; rather, it suggests that their dominant functional connectivity signatures are topologically distinct.

The limitations of this study outline directions for future research. First, the current implementation relied on a single type of rsFC feature (Pearson's correlation). As demonstrated by [53, 54], rsFC features can take various forms (e.g., tangent-based measures), each potentially highlighting different biomarkers. Second, our current workflow only supports static embedding spaces, yet rsFC features fluctuate due to dynamic neuronal activity [55]. Developing a dynamic version of this workflow to capture temporal embedding shifts is under consideration. Finally, while we performed cross-dataset validation, the study was restricted to two small scale datasets. To address this, our group is actively constructing a unified, multi-center, multi-disorder database to enable rigorous large-scale external validation. External validation on this standardized database is a primary objective of our immediate follow-up work.

Methods

To infer the relationships between disorders in the rsFC-informed embedding space, we propose the following rsFC-based embedding-relation workflow, which consists of three components: (1) the estimation of rsFC features for multiple neuropsychiatric disorders; (2) the mutual information-based embedding framework; and (3) the quantification of the neurobiological relationships between disorders (Fig. 7).

Estimation of rsFC features of disorders

The workflow begins by estimating the rsFC features of disorders from curated (or recorded) resting-state fMRI data. Given n subjects in the resting-state fMRI data, we defined p brain regions using the selected reference atlas to extract each subject's region-wise fMRI time courses. Next, we derived a square symmetric affinity matrix $\mathcal{A} \in \mathbb{R}^{p \times p}$ to represent the subject-wise rsFC features. Since different reference atlases and rsFC quantification measures yield varying rsFC features [53], we configured the reference atlas as the BSA-AAL parcellation atlas [36] and used the standard Pearson's r correlation [12] (with Fisher's transformation applied to the correlation values) for rsFC quantification. The BSA-AAL atlas was chosen for its comprehensive whole-brain coverage (140 ROIs) and to maintain consistency with prior multi-disorder studies [15]. As the obtained affinity matrix $\mathcal{A}$ is square and symmetric for each subject, we

vectorized each $\mathcal{A}$ by retaining its flattened lower triangular part. This reshapes the subject-wise affinity matrix into arrays representing the rsFC features: $\mathcal{A} \rightarrow X \in \mathbb{R}^L; L = \frac{p\cdot(p-1)}{2}$.

Mutual information-based embedding framework

The core of our workflow is the proposed mutual information-based embedding framework. Using the previously estimated rsFC features for n subjects, $x_1, \ldots, x_n \in \mathbb{R}^{n\times L}$, which exist in the L-dimensional feature space $\mathcal{U}$, we aim to learn low-dimensional embeddings $z_1, \ldots, z_n \in \mathbb{R}^{n\times l}$ that reside in the l-dimensional embedding space $\mathcal{V}$, where $L \gg l$. These low-dimensional embeddings represent the high-dimensional rsFC features. This embedding process can be described by the posterior distribution $p(z|x)$. To achieve the optimal $p(z|x)$, we aim to maximize the mutual information between the high-dimensional rsFC features and their corresponding low-dimensional embeddings as follows:

$$\mathcal{I}(x, z). \tag{1}$$

Depending on the imposed prior for embeddings, whether to embed subject-wise similarity in the embedding space, and the varying amount of added diagnostic information (derived from distinct clinical phenotypes), a series of embedding approaches are proposed under the mutual information-based embedding framework.

With differentiated imposed priors. To control the shape of the embedding space and simplify the cumbersome optimization of the naive mutual information term $\mathcal{I}(x, z)$ [56], we incorporate a prior distribution for the embeddings $q(z)$. This encourages the learned embedding posterior $p(z|x)$ to align closely with the imposed prior $q(z)$ by minimizing their density discrepancy:

$$\mathcal{I}(x, z) - \mathcal{D}_{KL}(p(z|x)||q(z)), \tag{2}$$

where $\mathcal{D}_{KL}$ denotes the K-L divergence in measuring the density difference between the posterior and imposed priors. For clarity, Eq. 2 can be viewed as a trade-off: maximizing $\mathcal{I}(x, z)$ while using the $\mathcal{D}_{KL}$ term as a regularizer to keep $p(z|x)$ close to the prior $q(z)$. This ensures the embedding retains most information from x while its distribution remains shaped by the chosen prior. With the derivation (SI §1.1.), the foregoing mutual information based optimization function $\mathcal{I}(x, z) - \mathcal{D}_{KL}(p(z|x)||q(z))$ can be rewritten into a more easy-to-compute format:

$$\mathbb{E}_{x\sim p(x)}\big[\mathbb{E}_{z\sim p(z|x)}[-\log p(x|z)] + \mathcal{D}_{KL}(p(z|x)||q(z))\big], \tag{3}$$

where the derived $p(x|z)$ can be regarded as a decoder in the auto-encoder context.

To examine the effect of priors on learning embeddings, we use two types of priors: a Gaussian prior and a von-Mises Fisher (vMF) prior, resulting in the `Gaussian` (Fig. 7①) and `vMF` embedding approaches (Fig. 7②) as follows:

- Gaussian prior
  Due to its computational efficiency and widespread use in statistical learning [56], a Gaussian prior is the default choice for learning low-dimensional Gaussian embeddings of FC features. Assuming the Gaussian prior $q(z)$ follows a standard normal

distribution, $q(z) = \mathcal{N}(0, 1)$, the posterior $p(z|x)$ can also be defined as a Gaussian, $p(z|x) = \mathcal{N}(\mu, \sigma)$, where the mean and scale parameters are parameterized by deep neural networks (SI §1.1.1.).

- von-Mises Fisher(vMF) prior
  In addition to the common choice of a Gaussian prior, we propose using another tractable density prior: the von-Mises Fisher (vMF) prior, to encourage the learning of a spherical low-dimensional embedding space ($\mathcal{S}^1$). Unlike a Gaussian prior, which may concentrate the embeddings around the origin [57], a vMF prior promotes uniformly distributed embeddings throughout the space, avoiding concentration in any specific position. Denoting κ as the concentration parameter (set to $\kappa = 20$) and the normalization factor $\mathcal{C}_{1,\kappa} = \{\frac{1}{\int_{\mathcal{S}^1} \exp(\mu^\intercal z)\mathcal{S}^1}\}$, the vMF prior for a two-dimensional embedding z, encouraging uniform distribution in the embedding space, can be defined as $q(z) = \mathcal{C}_{(1,\kappa=0)}$. The resulting posterior can then be defined as another vMF distribution: $p(z|x) = \mathcal{C}_{1,\kappa} \exp \kappa(\mu^\intercal z)$, where the learnable parameter μ can be parameterized by a deep neural network (SI §1.1.2.).

With embedded subject-wise similarity. In addition to controlling the shape of the embedding space, the ability to represent subject-wise similarity in the embedding space is also a valuable property for learning the optimal $p(z|x)$. Similar to the multidimensional scaling (MDS) [58], assume the pairwise dissimilarity of rsFC features between two subjects (samples) in the feature space is captured in a squared proximity matrix d^2x_{ij}, defined using the Euclidean distance to represent dissimilarity: $d^2x_{ij} = ||x_i - x_j||^2$. We aim to preserve this pairwise dissimilarity in the embedding space, i.e., $||z_i - z_j||^2$, where $z_i, z_j \sim p(z|x)$, by introducing an additional penalty term based on the squared error of the distance discrepancy to the original optimization function:

$$\mathcal{I}(x, z) - \mathcal{D}_{KL}(p(z|x)||q(z)) - \sum_{i \neq j}(d^2x_{ij} - ||z_i - z_j||^2)^2. \tag{4}$$

By explicitly embedding sample-wise similarity, we ensure that the pairwise distances between samples in the original feature space are preserved within the learned low-dimensional manifold. We designate this as the `distance-preserving` embedding (Fig. 7③). Notably, this approach is versatile and can be established upon either Gaussian or spherical (vMF) embedding priors.

With varying amount of added diagnostic information. Diagnostic information, which refers to the clinical phenotypes of disorder(s), is crucial for learning optimal embeddings of disorder rsFC features. However, this information is often scarce and challenging to collect in large-scale clinical trials. To address this, we developed three embedding approaches that accommodate varying amount of diagnostic information.

Theoretically, the available diagnostic information can be considered as discrete categorical variables $c \in C$ representing N diagnostic categories. We formalize this through a categorical prior distribution $q(c) = \text{Cat}(c|\pi)$ with parameter vector $\pi \in \mathbb{R}_+^N$ representing class probabilities. With the inclusion of diagnostic information, the previous mutual information-based embedding framework evolves into an interaction

information-based embedding framework, incorporating high-dimensional rsFC features, low-dimensional embeddings, and the diagnostic information, represented as $\mathcal{I}(x, z, c)$ (SI §1.2). In conjunction with the imposed priors and embedded feature-wise similarity, the optimization term can be written as:

$$\mathcal{I}(x, z, c) - \mathcal{D}_{KL}(p(z|x)||q(z)) - \sum_{i \neq j}(d^2 x_{ij} - ||z_i - z_j||^2)^{1/2}. \tag{5}$$

Regardless of the imposed prior, the effect of added diagnostic information can theoretically be reduced to the term $\sum_c q(c|z) \log \frac{p(z|x)}{q(z|c)}$, which can be further simplified into a computable form: $\frac{1}{2}||z - \mu_c||^2$ (SI §1.2.1. and SI §1.2.2.), where μ_c represents the cluster means of clinical phenotypes in the embedding space.

- With the least amount of diagnostic information
  With minimal involvement of diagnostic information, the value of μ_c can be obtained as the randomly sampled mean parameter of a parameterized Gaussian distribution, allowing the discrete variables C to remain unobservable. Consequently, the incorporated diagnostic information requires only the number of phenotypes, N, to learn the optimal embeddings. We refer to this method as the `least-diagnostic` embedding (Fig. 7④).
- With median amount of diagnostic information
  Notably, instead of using the randomly sampled μ_c in the previous `least-diagnostic` embedding the value of μ_c can also be obtained by precomputing the empirical means for N phenotype-wise clusters. The calculation of these empirical cluster means requires the discrete variables C to be observable, ensuring that the learned embeddings z are closer to the cluster centers of their corresponding clinical phenotypes. Since this method relies more heavily on diagnostic information than the previous one, we refer to it as the `median-diagnostic` embedding (Fig. 7⑤).
- With the most amount of diagnostic information
  To further leverage diagnostic information in modeling, in addition to using precomputed empirical phenotype means for μ_c, we adopt a supervised contrastive learning paradigm [59] to explicitly model diagnostic differences in the embedding space. We define a pair of latent vectors, $\overrightarrow{z_1}, \overrightarrow{z_2} \in Z$, to represent two encoded embeddings, a distance function D to measure their dissimilarity, and a binary class indicator $C = 0, 1$ to verify whether the latent vectors come from the same disorder phenotype. The supervised contrastive loss function can be expressed as:

$$(1 - C)\frac{1}{2}(D_{(\overrightarrow{z_1}, \overrightarrow{z_2})})^2 + C\frac{1}{2}\max(0, m - D_{(\overrightarrow{z_1}, \overrightarrow{z_2})})^2, \tag{6}$$

  where m is a pre-defined margin that weighs the contribution of dissimilar pairs, and we use the Euclidean distance to compute $D_{(\overrightarrow{z_1}, \overrightarrow{z_2})}$, i.e., $D_{(\overrightarrow{z_1}, \overrightarrow{z_2})} = ||\overrightarrow{z_1} - \overrightarrow{z_2}||_2$. To preserve the authentic diagnostic differences among the encoded latent variables, we set $m = 0$, giving equal importance to dissimilar and similar pairs. Incorporating

this loss function into the optimization of Eq. 5 led to the development of the `most-diagnostic` embedding (Fig. 7⑥).

For summarization, we can further interpret the derived embedding approaches as graphical models, which are inline with VAE and its variations [23] (Fig. 8). Without the added diagnostic information, the former three embedding approaches (① ② ③) can be expressed using an identical graphical model. The differences among these three graphical models depends on differentiated priors on z (① ②) and whether or not to add pair-wise difference between subjects (③). The later three graphic models correspond to the use of varying amount of diagnostic information in mentioned embedding approaches (④ ⑤ ⑥).

Quantification of relationships among disorders in the embedding space

Leveraging these embedding approaches, we quantified the relationships among disorders in the learned low-dimensional rsFC-informed embedding space. These relationships were inferred directly from the between-cluster distances. We utilized the Frechet Inception Distance (FID) [34] as our primary metric to evaluate these relationships. Unlike conventional point-to-point metrics (e.g., L1 or L2 distance between centroids), FID is advantageous because it accounts for the full distribution of the clusters, i.e, incorporating both the mean (μ) and covariance (Σ) structure:

$$\mathcal{D}_{\mathrm{FID}}(\mu_1, \Sigma_1, \mu_2, \Sigma_2) = \mid \mu_1 - \mu_2 \mid^2 + tr(\Sigma_1 + \Sigma_2 - 2(\Sigma_1\Sigma_2)^{1/2}). \tag{7}$$

Here, μ and Σ represent the means and variances of the learned rsFC embeddings of each corresponding disorder phenotypes. In the case of the three major disorders:ASD, MDD, and SCZ, their neurobiological relationships in a connectivity-informed embedding space can then be inferred from their pairwise distances in the embedding space, represented as $\mathcal{D}_{ASD-SCZ}$, $\mathcal{D}_{ASD-MDD}$, and $\mathcal{D}_{MDD-SCZ}$.

While our proposed method utilizes a vMF prior that induces a spherical embedding space, we prioritize FID for its sensitivity to cluster dispersion. To ensure that our conclusions are not artifacts of distance metric, we also computed the geodesic distance as a secondary metric. These results are provided in the Supplementary Information (SI §3 and 4).

Alternative embedding methods

We benchmarked our embedding methods against widely used linear, manifold-learning, graph-based, metric-learning, and deep generative embedding approaches applied only to the synthetic rsFC dataset. Specifically, we evaluated linear PCA, local linear embedding (LLE), Graph2Vec (Graph2V), t-SNE (with and without diagnostic supervision), UMAP (unsupervised, diagnostically supervised, and hyperbolic instantiations), neighborhood component analysis (NCA), and a fully supervised variational autoencoder (VAE). The resulting embeddings and comparative performance

on synthetic rsFC data are presented in Fig. 3. For reproducibility, full technical specifications (architectures where applicable, optimization settings, and hyperparameters) are provided in SI §2 and summarized in Table S3.

Declaration statements

- Data Availability
  The employed UTO and HuShoWa datasets are derived from the SRPBS database under the Brain/MINDS Beyond human brain MRI project [14], which is consisted of resting-state fMRI scans from 8 different sites. The data is openly accessible via https://bicr-resource.atr.jp/srpbsfc/.
- Code Availability
  The python code for executing our embedding-relation workflow on designated rsFC datasets can be found at https://github.com/LeonBai/rsFC_embedding.
- Acknowledgements
  We thank Dr. Ayumu Yamashita (The University of Tokyo) and Prof. Kenji Doya (OIST) for comments on an early version of the manuscript. This work was supported by the Japan Agency for Medical Research and Development (AMED) under Grant Numbers JP24dm0307008, JP25wm0625204s0102, and JP25wm0625122s0502. The funders had no role in the study design, data collection, analysis, and results interpretation.
- Authors' Contributions
  WJ.B. conceives the general idea, and performs the overall analysis while O.Y., Y.S., and J.Y. provide fruitful feedback and suggestions throughout the analysis and revision procedure. WJ.B. wrote the draft of the manuscript. All authors discussed the results and commented on the manuscript.
- Competing Interests
  The authors declare no competing financial or non-financial interests.

References

[1] Oldham, J. M. *et al.* Comorbidity of axis i and axis ii disorders. *The American Journal of Psychiatry* (1995).

[2] Angold, A., Costello, E. J. & Erkanli, A. Comorbidity. *The Journal of Child Psychology and Psychiatry and Allied Disciplines* **40**, 57–87 (1999).

[3] Helzer, J. E. *et al. Dimensional approaches in diagnostic classification: Refining the research agenda for DSM-V* (American Psychiatric Pub, 2009).

[4] Casey, B. *et al.* Dsm-5 and rdoc: progress in psychiatry research? *Nature Reviews Neuroscience* **14**, 810–814 (2013).

[5] Caspi, A. & Moffitt, T. E. All for one and one for all: Mental disorders in one dimension. *American Journal of Psychiatry* **175**, 831–844 (2018).

[6] Anttila, V. *et al.* Analysis of shared heritability in common disorders of the brain. *Science* **360** (2018).

[7] Gandal, M. J. *et al.* Shared molecular neuropathology across major psychiatric disorders parallels polygenic overlap. *Science* **359**, 693–697 (2018).

[8] Lee, P. H., Feng, Y.-C. A. & Smoller, J. W. Pleiotropy and cross-disorder genetics among psychiatric disorders. *Biological psychiatry* **89**, 20–31 (2021).

[9] Chen, Y., Li, W., Lv, L. & Yue, W. Shared genetic determinants of schizophrenia and autism spectrum disorder implicate opposite risk patterns: A genome-wide analysis of common variants. *Schizophrenia bulletin* **50**, 1382–1395 (2024).

[10] Abi-Dargham, A. & Horga, G. The search for imaging biomarkers in psychiatric disorders. *Nature medicine* **22**, 1248–1255 (2016).

[11] Woo, C.-W., Chang, L. J., Lindquist, M. A. & Wager, T. D. Building better biomarkers: brain models in translational neuroimaging. *Nature neuroscience* **20**, 365–377 (2017).

[12] Friston, K. J. Functional and effective connectivity: a review. *Brain connectivity* **1**, 13–36 (2011).

[13] Sudlow, C. *et al.* Uk biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS medicine* **12**, e1001779 (2015).

[14] Koike, S. *et al.* Brain/minds beyond human brain mri project: a protocol for multi-level harmonization across brain disorders throughout the lifespan. *NeuroImage: Clinical* 102600 (2021).

[15] Yahata, N. *et al.* A small number of abnormal brain connections predicts adult autism spectrum disorder. *Nature communications* **7**, 1–12 (2016).

[16] Ichikawa, N. *et al.* Primary functional brain connections associated with melancholic major depressive disorder and modulation by antidepressants. *Scientific reports* **10**, 1–12 (2020).

[17] Yamashita, A. *et al.* Generalizable brain network markers of major depressive disorder across multiple imaging sites. *PLoS biology* **18**, e3000966 (2020).

[18] Shen, H., Wang, L., Liu, Y. & Hu, D. Discriminative analysis of resting-state functional connectivity patterns of schizophrenia using low dimensional embedding of fmri. *Neuroimage* **49**, 3110–3121 (2010).

[19] Drysdale, A. T. *et al.* Resting-state connectivity biomarkers define neurophysiological subtypes of depression. *Nature medicine* **23**, 28–38 (2017).

[20] Tokuda, T. *et al.* Identification of depression subtypes and relevant brain regions using a data-driven approach. *Scientific reports* **8**, 1–13 (2018).

[21] Huys, Q. J., Maia, T. V. & Frank, M. J. Computational psychiatry as a bridge from neuroscience to clinical applications. *Nature neuroscience* **19**, 404–413 (2016).

[22] Aggarwal, C. C., Hinneburg, A. & Keim, D. A. On the surprising behavior of distance metrics in high dimensional space. International Conference on Database Theory (2001). Pp. 420–434.

[23] Kingma, D. P. & Welling, M. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114* (2013).

[24] Kingma, D. P., Rezende, D. J., Mohamed, S. & Welling, M. Semi-supervised learning with deep generative models. Advances in Neural Information Processing Systems (2014). Pp. 3581–3589.

[25] Li, X., Geenjaar, E., Fu, Z., Pearlson, G. D. & Calhoun, V. D. Brain functional network connectivity interpolation characterizes neuropsychiatric continuum and heterogeneity. *bioRxiv* 2024–11 (2024).

[26] Craddock, N. & Owen, M. J. The kraepelinian dichotomy–going, going but still not gone. *The British Journal of Psychiatry* **196**, 92–95 (2010).

[27] Tanaka, S. C. *et al.* A multi-site, multi-disorder resting-state magnetic resonance image database. *Scientific data* **8**, 1–15 (2021).

[28] Roweis, S. T. & Saul, L. K. Nonlinear dimensionality reduction by locally linear embedding. *science* **290**, 2323–2326 (2000).

[29] Narayanan, A. *et al.* graph2vec: Learning distributed representations of graphs. *arXiv preprint arXiv:1707.05005* (2017).

[30] Van der Maaten, L. & Hinton, G. Visualizing data using t-sne. *Journal of machine learning research* **9** (2008).

[31] McInnes, L., Healy, J. & Melville, J. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426* (2018).

[32] Goldberger, J., Hinton, G. E., Roweis, S. & Salakhutdinov, R. R. Neighbourhood components analysis. *Advances in neural information processing systems* **17**, 513–520 (2004).

[33] Khemakhem, I., Kingma, D. P., Monti, R. & Hyvärinen, A. Variational autoencoders and nonlinear ICA: A unifying framework. International Conference on Artificial Intelligence and Statistics (2020). PMLR, pp. 2207–2217.

[34] Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B. & Hochreiter, S. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *arXiv preprint arXiv:1706.08500* (2017).

[35] Zeng, W. *et al.* Gmaplatent: Geometric mapping in latent space. *arXiv preprint arXiv:2503.23407* (2025).

[36] Perrot, M., Rivière, D. & Mangin, J.-F. Cortical sulci recognition and spatial normalization. *Medical image analysis* **15**, 529–550 (2011).

[37] Yu, M. *et al.* Statistical harmonization corrects site effects in functional connectivity measurements from multi-site fmri data. *Human brain mapping* **39**, 4213–4227 (2018).

[38] Yamashita, O. *et al.* Computational mechanisms of neuroimaging biomarkers uncovered by multicenter resting-state fmri connectivity variation profile. *Molecular Psychiatry* **30**, 5463–5474 (2025).

[39] Montavon, G., Binder, A., Lapuschkin, S., Samek, W. & Müller, K.-R. Layer-wise relevance propagation: an overview. *Explainable AI: interpreting, explaining and visualizing deep learning* 193–209 (2019).

[40] Yeo, B. T. *et al.* The organization of the human cerebral cortex estimated by intrinsic functional connectivity. *Journal of neurophysiology* (2011).

[41] Baron-Cohen, S., Wheelwright, S., Skinner, R., Martin, J. & Clubley, E. The autism-spectrum quotient (aq): Evidence from asperger syndrome/high-functioning autism, malesand females, scientists and mathematicians. *Journal of autism and developmental disorders* **31**, 5–17 (2001).

[42] Beck, A. T., Steer, R. A., Ball, R. & Ranieri, W. F. Comparison of beck depression inventories-ia and-ii in psychiatric outpatients. *Journal of personality assessment* **67**, 588–597 (1996).

[43] Vos de Wael, R. *et al.* Brainspace: a toolbox for the analysis of macroscale gradients in neuroimaging and connectomics datasets. *Communications biology* **3**, 1–10 (2020).

[44] Morrissey, Z. D., Zhan, L., Ajilore, O. & Leow, A. D. rest2vec: Vectorizing the resting-state functional connectome using graph embedding. *NeuroImage* **226**, 117538 (2021).

[45] Whi, W. *et al.* Characteristic functional cores revealed by hyperbolic disc embedding and k-core percolation on resting-state fmri. *Scientific reports* **12**, 1–16 (2022).

[46] Espadoto, M., Martins, R. M., Kerren, A., Hirata, N. S. & Telea, A. C. Toward a quantitative survey of dimension reduction techniques. *IEEE transactions on visualization and computer graphics* **27**, 2153–2173 (2019).

[47] Nordgaard, J., Sass, L. A. & Parnas, J. The psychiatric interview: validity, structure, and subjectivity. *European archives of psychiatry and clinical neuroscience* **263**, 353–364 (2013).

[48] Razafsha, M. *et al.* Biomarker identification in psychiatric disorders: from neuroscience to clinical practice. *Journal of Psychiatric Practice®* **21**, 37–48 (2015).

[49] Hettwer, M. *et al.* Coordinated cortical thickness alterations across six neurodevelopmental and psychiatric disorders. *Nature communications* **13**, 6851 (2022).

[50] Qi, S. *et al.* Common and unique multimodal covarying patterns in autism spectrum disorder subtypes. *Molecular autism* **11**, 90 (2020).

[51] Sadeghi, I. *et al.* Brain transcriptomic profiling reveals common alterations across neurodegenerative and psychiatric disorders. *Computational and Structural Biotechnology Journal* **20**, 4549–4561 (2022).

[52] Chien, Y.-L., Wu, C.-S. & Tsai, H.-J. The comorbidity of schizophrenia spectrum and mood disorders in autism spectrum disorder. *Autism Research* **14**, 571–581 (2021).

[53] Dadi, K. *et al.* Benchmarking functional connectome-based predictive models for resting-state fmri. *NeuroImage* **192**, 115–134 (2019).

[54] Takahara, Y. *et al.* Comprehensive evaluation of pipelines for classification of psychiatric disorders using multi-site resting-state fmri datasets. *Neural Networks* **187**, 107335 (2025).

[55] Preti, M. G., Bolton, T. A. & Van De Ville, D. The dynamic functional connectome: State-of-the-art and perspectives. *Neuroimage* **160**, 41–54 (2017).

[56] Murphy, K. P. *Machine learning: a probabilistic perspective* (MIT press, 2012).

[57] Davidson, T. R., Falorsi, L., De Cao, N., Kipf, T. & Tomczak, J. M. Hyperspherical variational auto-encoders. Conference on Uncertainty in Artificial Intelligence (2018). Pp. 856–865.

[58] Saez, M. Modern multidimensional scaling: Theory and applications. (1998).

[59] Hadsell, R., Chopra, S. & LeCun, Y. Dimensionality reduction by learning an invariant mapping. IEEE Conference on Computer Vision and Pattern Recognition (2006). Pp. 1735–1742.

Figures, Tables and associated legends

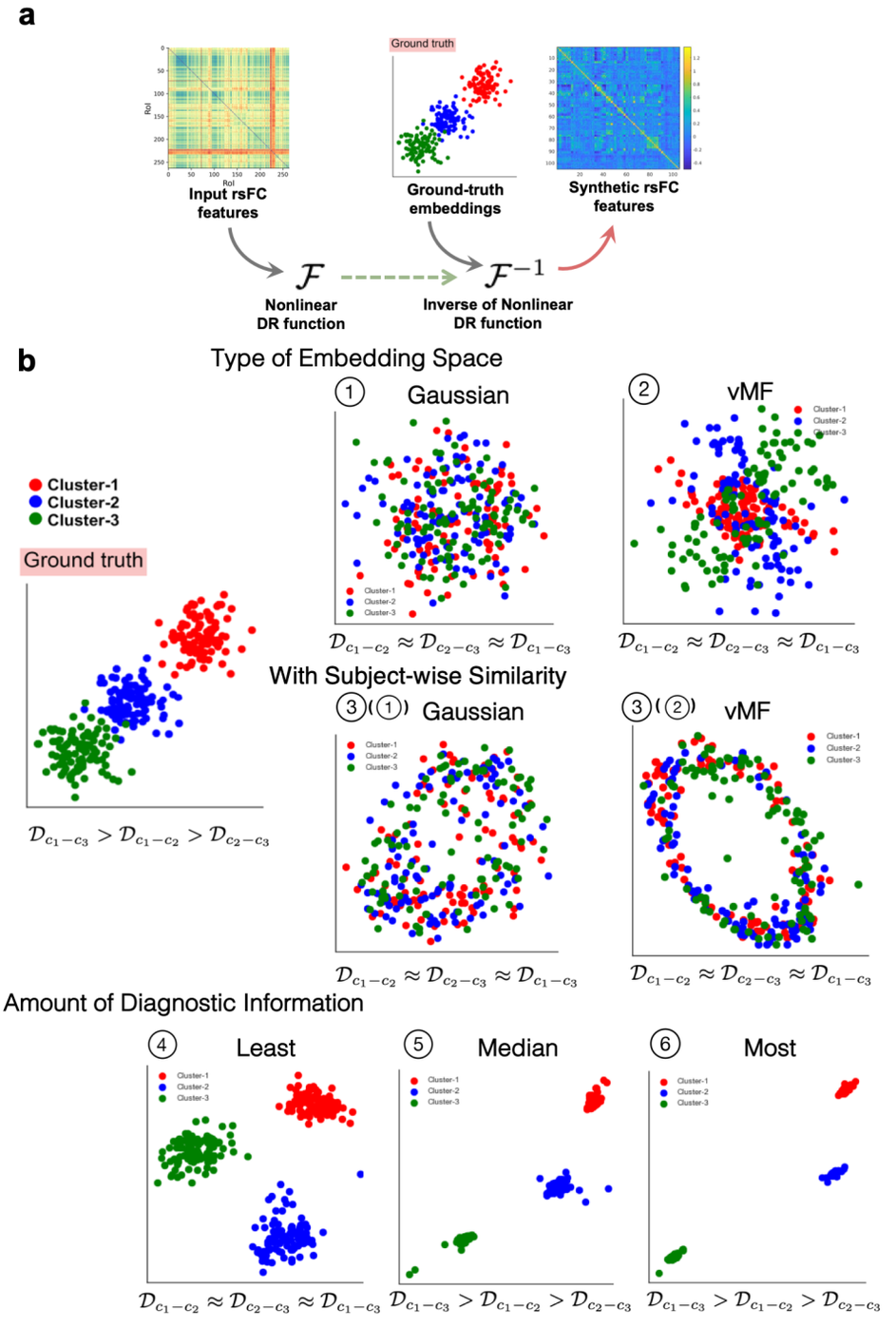
a
Ground truth
Input rsFC features
Ground-truth embeddings
Synthetic rsFC features
$\mathcal{F}$
$\mathcal{F}^{-1}$
Nonlinear DR function
Inverse of Nonlinear DR function
b
Type of Embedding Space
Cluster-1
Cluster-2
Cluster-3
Ground truth
$\mathcal{D}_{c_1-c_3} > \mathcal{D}_{c_1-c_2} > \mathcal{D}_{c_2-c_3}$
① Gaussian
$\mathcal{D}_{c_1-c_2} \approx \mathcal{D}_{c_2-c_3} \approx \mathcal{D}_{c_1-c_3}$
② vMF
$\mathcal{D}_{c_1-c_2} \approx \mathcal{D}_{c_2-c_3} \approx \mathcal{D}_{c_1-c_3}$
With Subject-wise Similarity
③(①) Gaussian
$\mathcal{D}_{c_1-c_2} \approx \mathcal{D}_{c_2-c_3} \approx \mathcal{D}_{c_1-c_3}$
③(②) vMF
$\mathcal{D}_{c_1-c_2} \approx \mathcal{D}_{c_2-c_3} \approx \mathcal{D}_{c_1-c_3}$
Amount of Diagnostic Information
④ Least
$\mathcal{D}_{c_1-c_2} \approx \mathcal{D}_{c_2-c_3} \approx \mathcal{D}_{c_1-c_3}$
⑤ Median
$\mathcal{D}_{c_1-c_3} > \mathcal{D}_{c_1-c_2} > \mathcal{D}_{c_2-c_3}$
⑥ Most
$\mathcal{D}_{c_1-c_3} > \mathcal{D}_{c_1-c_2} > \mathcal{D}_{c_2-c_3}$

Fig. 1 Evaluation of embedding approaches derived from the mutual information-based framework on synthetic rsFC data.
(a) Synthetic rsFC features generation process. To generate synthetic features corresponding to ground-truth low-dimensional embeddings, a nonlinear dimensionality reduction (DR) function $\mathcal{F}$ was first trained on rsFC features from an open public dataset. Ground-truth embeddings were then generated with defined pairwise distances and passed through the inverse of the learned projection function ($\mathcal{F}^{-1}$) to produce synthetic high-dimensional rsFC features matching the dimensionality of the input data.
(b) Performance assessment of embedding approaches. The learned embedding spaces are visualized alongside the statistically verified inter-cluster relationships (determined via paired t-tests with Bonferroni correction over 10-fold cross-validation; see SI §3 and Fig. S1). Abbreviations: C1(red-colored)/C2(blue-colored)/C3(green-colored) denote Cluster 1/2/3.

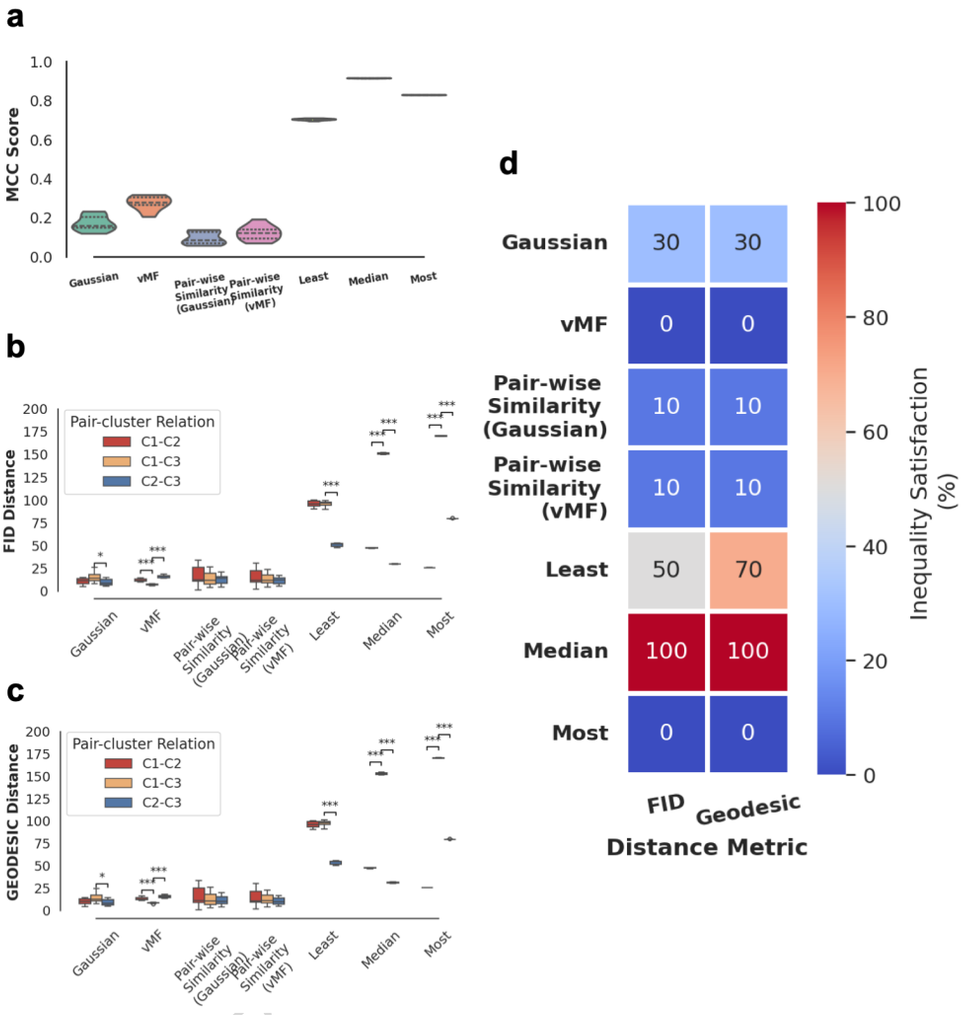

Fig. 2 Evaluation of embedding quality and topological preservation on synthetic rsFC data.
(a) Mean Correlation Coefficient (MCC) scores across the employed embedding approaches, demonstrating their respective efficacy in capturing the original data structure. The proposed method, utilizing median diagnostic information, achieves the highest performance.
(b, c) Assessment of structural (between-cluster relationships) preservation via (b) FID and (c) Geodesic distance metrics. The plots reveal how distinct embedding approaches reconstruct the relative distances between clusters (C1, C2, C3).
(d) Evaluation of topological consistency. The heatmap reports the percentage of satisfaction for the ground-truth inequality relation ($\mathcal{D}_{c_1-c_3} > \mathcal{D}_{c_1-c_2} > \mathcal{D}_{c_2-c_3}$). Notably, the embedding method with the median amount of diagnostic information achieves 100% satisfaction, indicating robust recovery of the underlying manifold structure across both distance metrics.

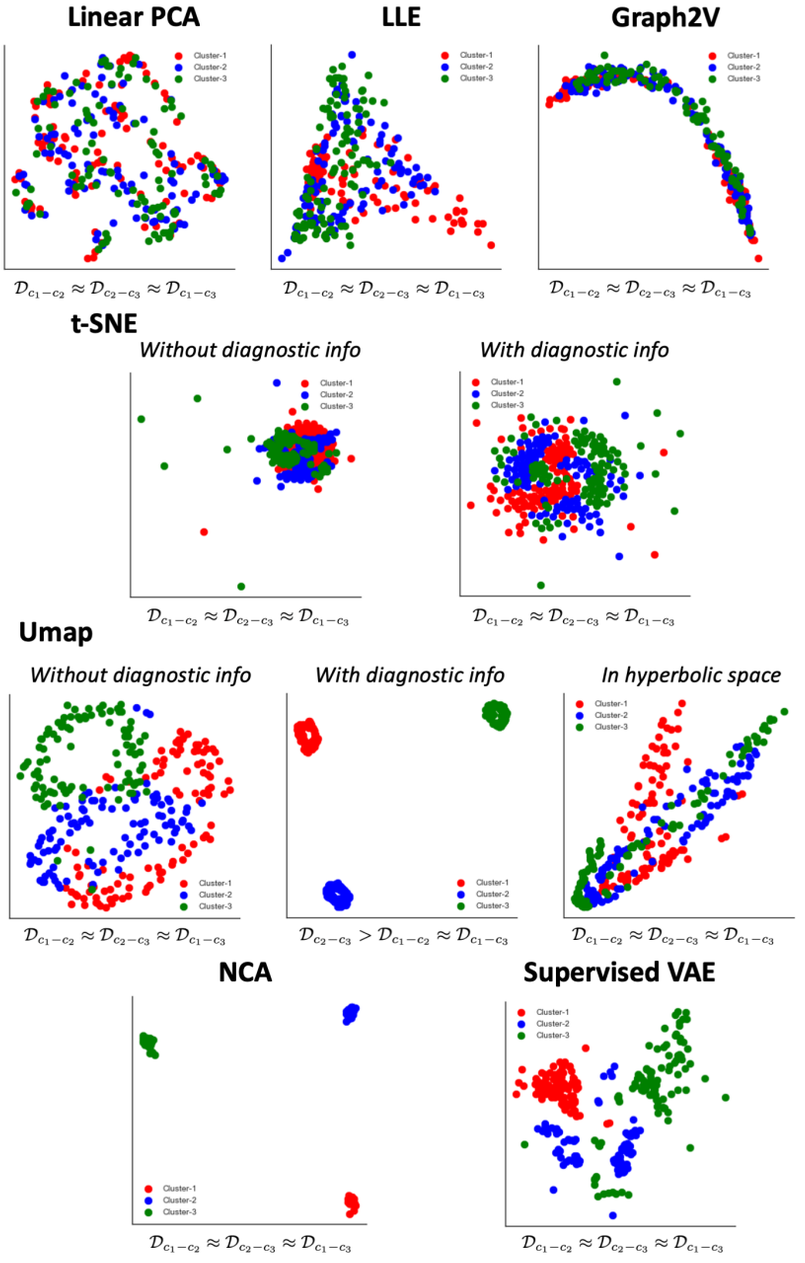

Fig. 3 Evaluation of alternative embedding approaches on synthetic rsFC data. We evaluate the performance of ten alternative embedding approaches in recovering the ground-truth between-cluster relationships ($\mathcal{D}_{c_1-c_3} > \mathcal{D}_{c_1-c_2} > \mathcal{D}_{c_2-c_3}$). Evaluated embedding methods include linear projection (PCA), manifold learning (LLE, Graph2V), unsupervised non-linear embeddings (t-SNE, UMAP without diagnostic info), and supervised metric learning (NCA, Supervised VAE, UMAP).

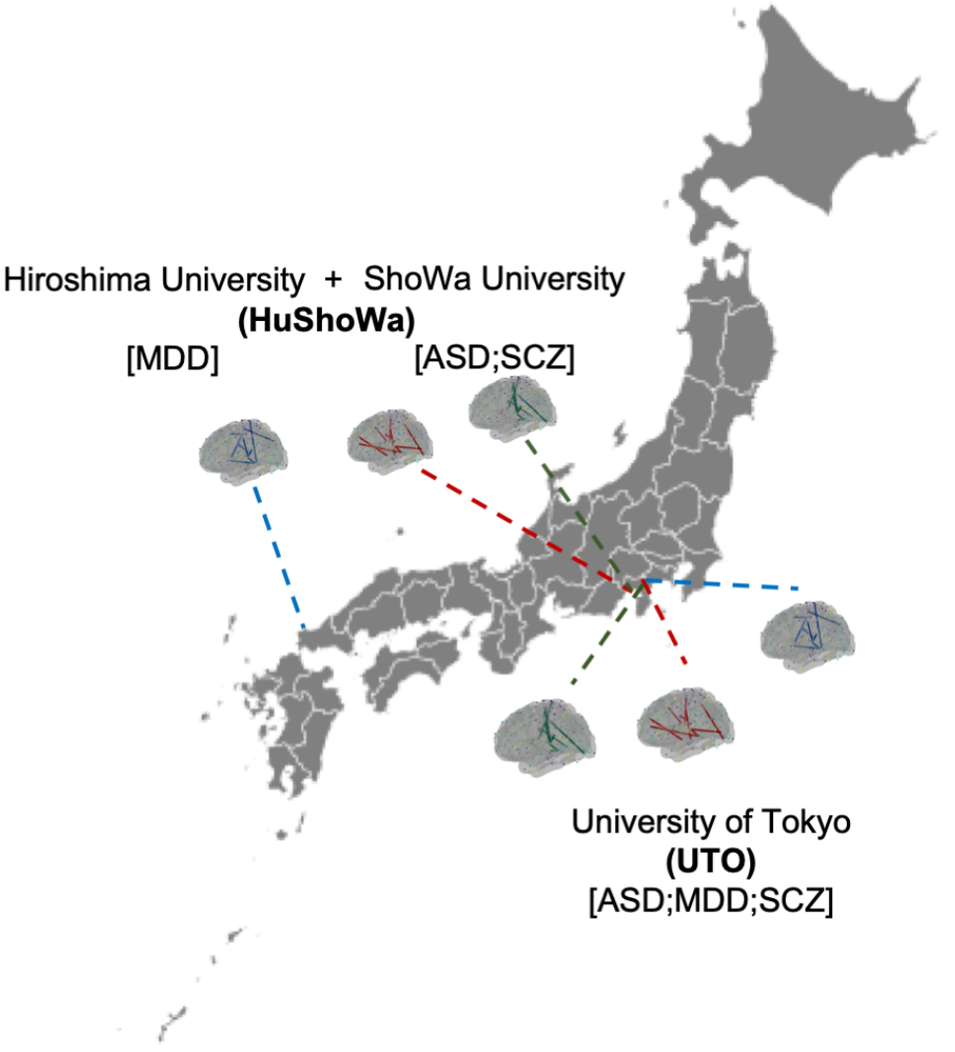

Fig. 4 Geographic information for the curated UTO and HuShoWa datasets.
In the UTO dataset, participants with one of the main disorders (ASD, MDD, and SCZ) were scanned at a single site (University of Tokyo, Tokyo, Japan). In the HuShoWa dataset, participants with MDD were scanned at Hiroshima University (Hiroshima, Japan), whereas participants with ASD or SCZ were scanned at Showa University (Tokyo, Japan). The layout and styling of the site-location map are adapted from the visualization in Fig. 2 of [14].

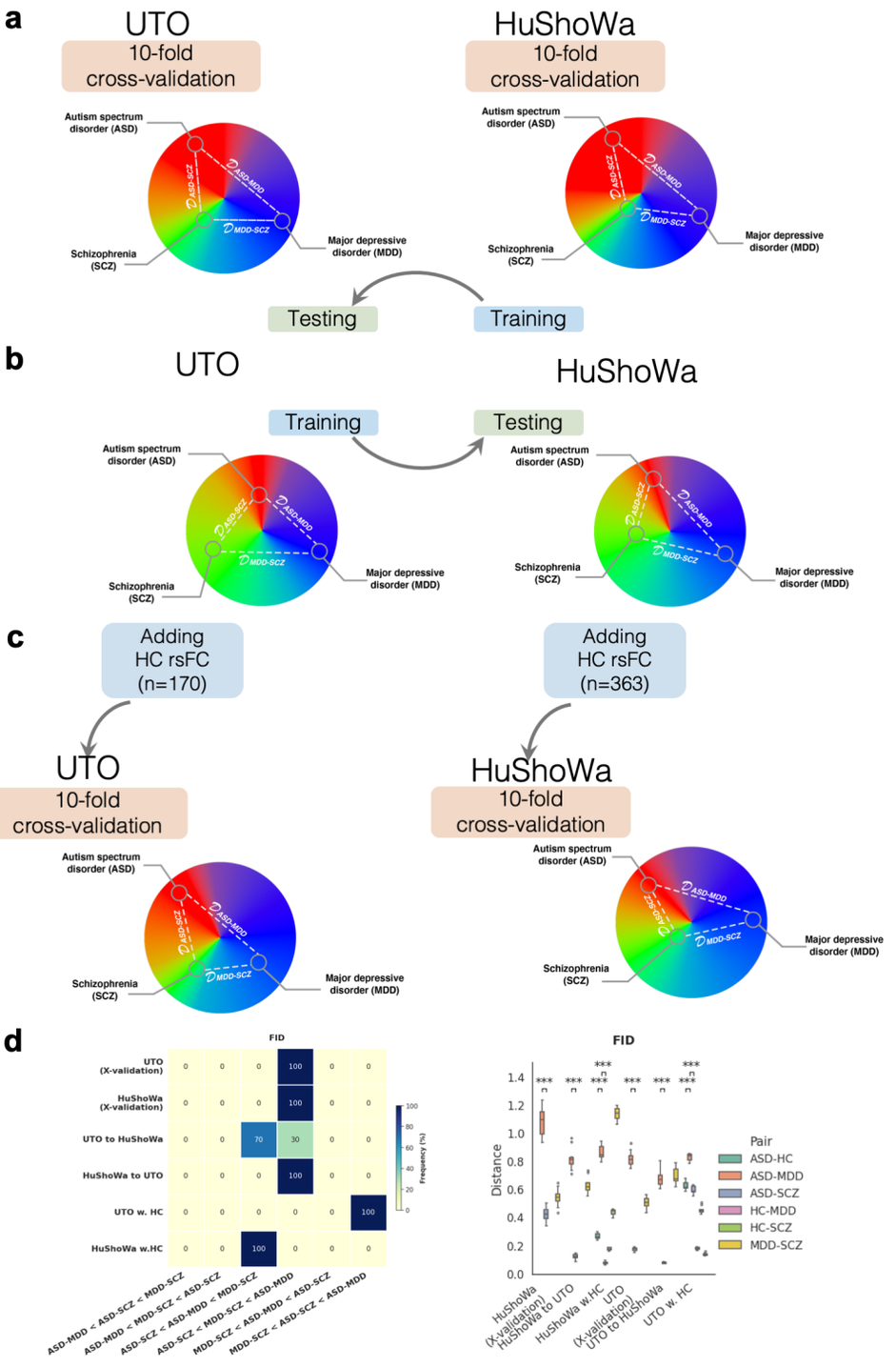
a
UTO
10-fold cross-validation
HuShoWa
10-fold cross-validation
Autism spectrum disorder (ASD)
Schizophrenia (SCZ)
Major depressive disorder (MDD)
ASD-SCZ
ASD-MDD
MDD-SCZ
Testing
Training
b
UTO
HuShoWa
Training
Testing
c
Adding HC rsFC (n=170)
Adding HC rsFC (n=363)
UTO
10-fold cross-validation
HuShoWa
10-fold cross-validation
d
FID
UTO (X-validation)
HuShoWa (X-validation)
UTO to HuShoWa
HuShoWa to UTO
UTO w. HC
HuShoWa w.HC
Frequency (%)
ASD-MDD < ASD-SCZ < MDD-SCZ
ASD-MDD < MDD-SCZ < ASD-SCZ
ASD-SCZ < ASD-MDD < MDD-SCZ
ASD-SCZ < MDD-SCZ < ASD-MDD
MDD-SCZ < ASD-MDD < ASD-SCZ
MDD-SCZ < ASD-SCZ < ASD-MDD
FID
Distance
Pair
ASD-HC
ASD-MDD
ASD-SCZ
HC-MDD
HC-SCZ
MDD-SCZ

Fig. 5 Neurobiological relationship between ASD, MDD and SCZ in rsFC-informed embedding spaces. Evaluations were performed on the curated UTO and HuShoWa datasets using three validation schemes: (a) Conventional 10-fold cross-validation; (b) Cross-dataset validation (UTO vs. HuShoWa); and (c) Validation incorporating rsFC features from Healthy Control (HC) participants.
In (a-c), the left panels visualize the embedding space (axes represent the first two dimensions), while the right panels display the computed inter-disorder distances. Raw quantitative results are provided in Fig. S3.
(d) Consistency of neurobiological structures. Heatmaps display the frequency of specific distance ranking patterns (inequality relations) based on FID scores (left) and raw feature results (right). Darker colors indicate a higher recurrence of a specific ordering (e.g., $D_{MDD\text{-}SCZ} < D_{ASD\text{-}SCZ} < D_{ASD\text{-}MDD}$) across 10 folds. Results for the secondary geodesic-based metric are provided in Fig. S4, detailed embedding visualizations are available in Fig. S3, and permutation test results for UTO dataset are also reported in Fig. S7.

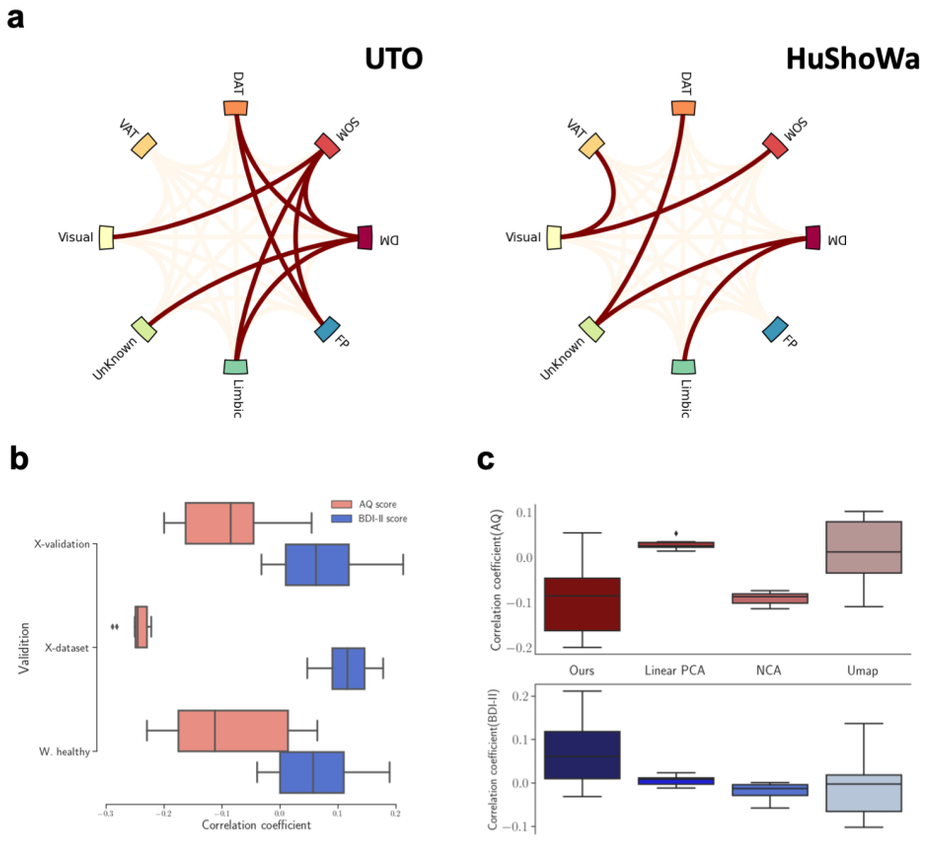

Fig. 6 Neuroscientific interpretation of the rsFC-informed embedding space.
(a) The top contributing inter-regional rsFC pairs from a network perspective for the UTO and HuShoWa datasets. For clarity, within-network connections are omitted. Detailed region names and laterality are listed in Table S5 and S6. Network abbreviations: DAT: dorsal attention; DM: default-mode; SOM: somatomotor; FP: fronto-parietal.
(b) Clinical relevance of the learned embedding space (HuShoWa dataset). Scatter plots show the correlation between composite embedding scores (y-axis) and clinical severity scores (x-axis; Autism Spectrum Quotient (AQ score) for ASD and Beck Depression Inventory (BDI score) for MDD) across three validation schemes: 10-fold cross-validation (X-validation), cross-dataset validation (X-dataset), and healthy-control validation (W.healthy).
(c) Comparative analysis of clinical correlations. The y-axis represents the correlation coefficients between embedding-derived scores and clinical severity, compared across our method and several alternative embedding approaches (x-axis) under standard 10-fold cross-validation.

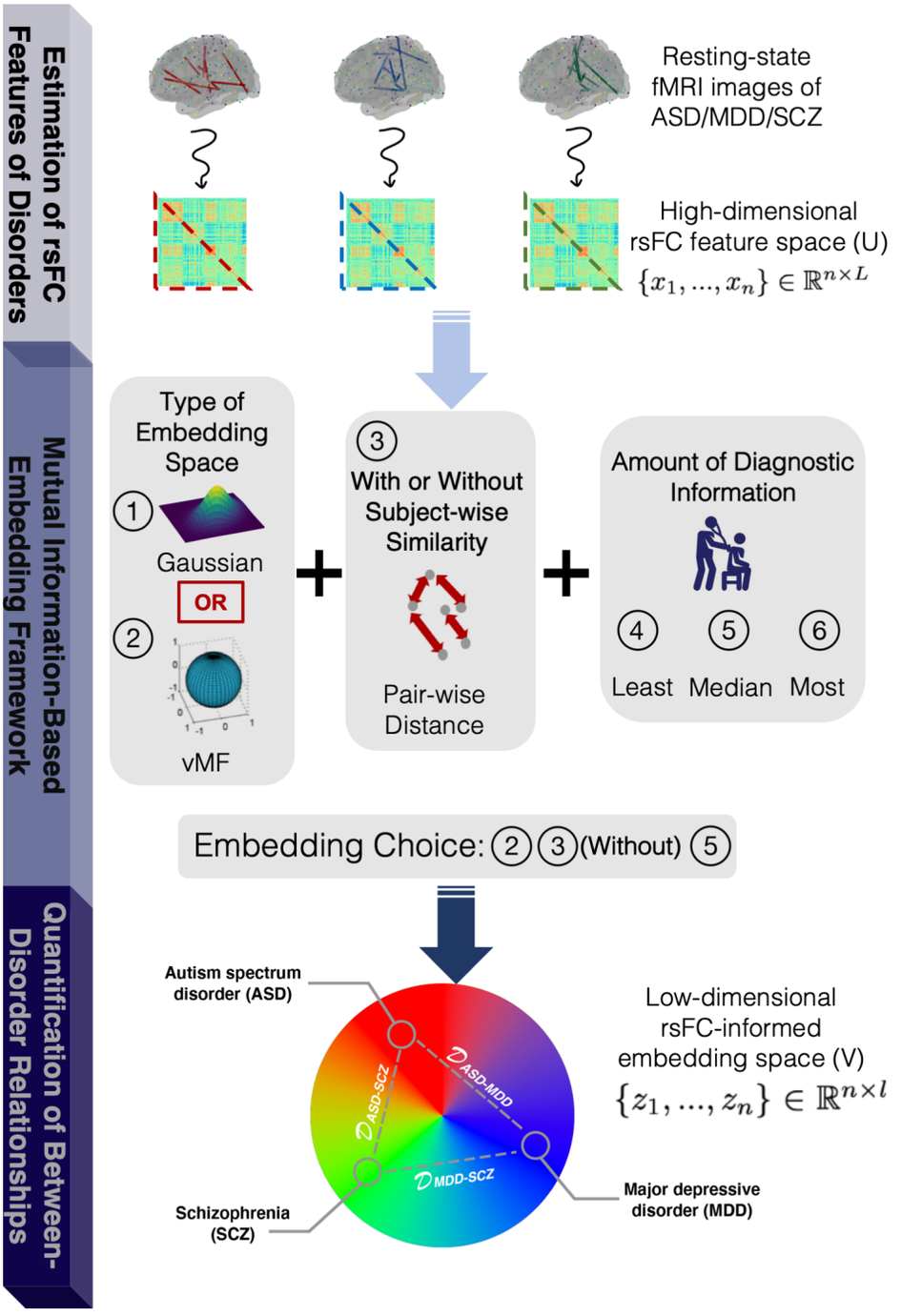
Estimation of rsFC Features of Disorders
Mutual Information-Based Embedding Framework
Quantification of Between-Disorder Relationships
Resting-state fMRI images of ASD/MDD/SCZ
High-dimensional rsFC feature space (U)
$\{x_1, ..., x_n\} \in \mathbb{R}^{n \times L}$
Type of Embedding Space
1
Gaussian
OR
2
vMF
3
With or Without Subject-wise Similarity
Pair-wise Distance
Amount of Diagnostic Information
4
Least
5
Median
6
Most
Embedding Choice: 2 3 (Without) 5
Autism spectrum disorder (ASD)
Schizophrenia (SCZ)
Major depressive disorder (MDD)
$\mathcal{D}_{ASD\text{-}MDD}$
$\mathcal{D}_{ASD\text{-}SCZ}$
$\mathcal{D}_{MDD\text{-}SCZ}$
Low-dimensional rsFC-informed embedding space (V)
$\{z_1, ..., z_n\} \in \mathbb{R}^{n \times l}$

Fig. 7 The diagram on the proposed rsFC-based embedding-relation workflow.
(1) Estimating rsFC features from resting-state fMRI images of multiple disorders. We derive a square affinity matrix and further vectorize it as 1D arrays (dotted line expressed lower triangular part) to represent rsFC features for each subject. These rsFC features live in the high-dimensional rsFC feature space $\mathcal{U}$.
(2) Mutual information-based embedding framework. The primary goal is to maximize the mutual information between rsFC features x in the high-dimensional feature space $\mathcal{U}$ and their low-dimensional embeddings z in the embedding space $\mathcal{V}$, i.e., $\mathcal{I}(x, z)$. ① and ②: Depending on the imposed priors in controlling the shape of embedding space, two embedding approaches in opting for Gaussian and vMF priors are presented here. ③: Allowing the pair-wise difference between embeddings to reflect their subject-wise dissimilarity on the feature space, we put forward the distance- preserving embedding approach to embed the pair-wise distance between subjects. ④, ⑤, ⑥: Depending on the varying amount of diagnostic information in use, we further develop three embedding approaches that correspond to the utilization of the least/median/most amount of diagnostic information in learning optimal embeddings of rsFC features. Note here, with added diagnostic information, these three embedding methods are framed based on an interaction information-based embedding framework, i.e., $\mathcal{I}(x, z, c)$.
(3) Quantification of relationships between disorders on the embedding space. Once an optimal embedding space that represents the rsFC features of all disorders (e.g., ASD, MDD, SCZ) has been learned, the relationship between any two disorders can be quantified by their pair-wise distance in this embedding space (e.g., $\mathcal{D}_{ASD-SCZ}$, $\mathcal{D}_{ASD-MDD}$, $\mathcal{D}_{MDD-SCZ}$).

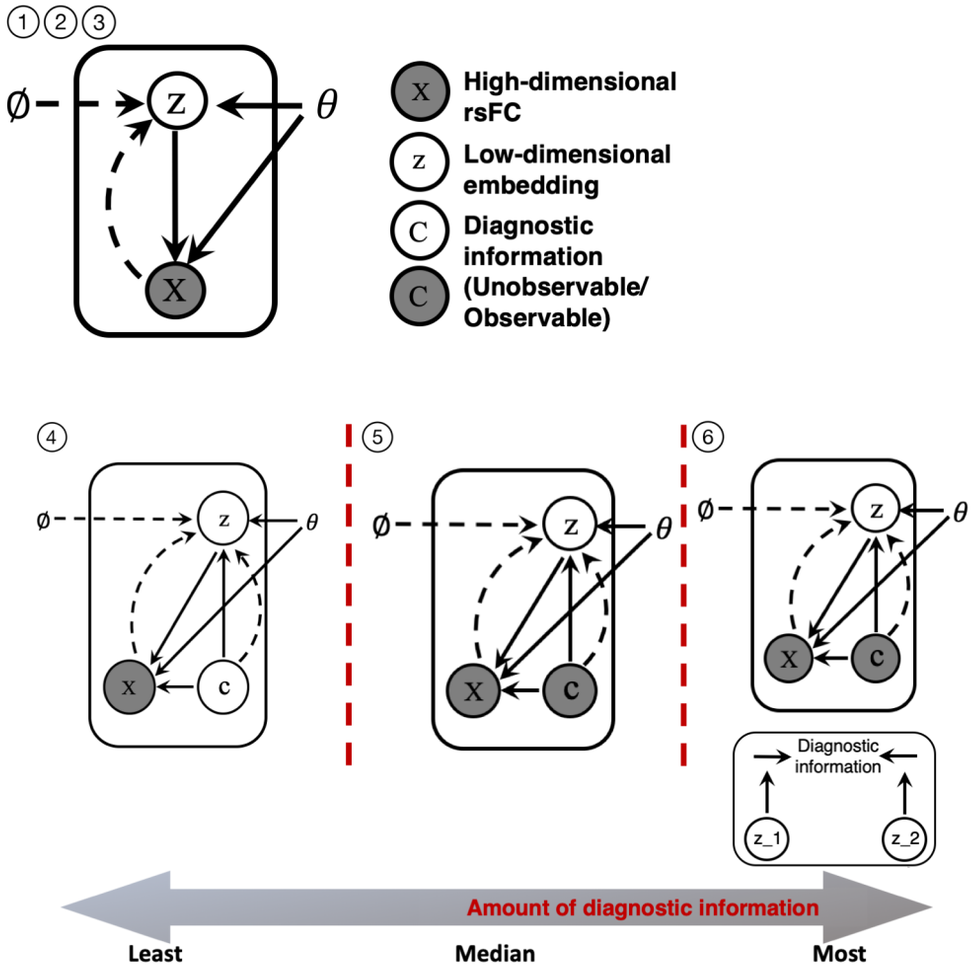

Fig. 8 Graphical model representations of the proposed embedding approaches. Across the proposed embedding approaches, x represents the high-dimensional rsFC features obtained from the first component of our workflow. The model aims to learn the low-dimensional latent variable z, optionally utilizing diagnostic information represented by the discrete variable c. Consistent with standard VAE terminology [23], solid lines denote generative processes, while dashed lines indicate inference processes.

Supplementary information

Supplementary Information (§1–§5; Figures S1–S9; Tables S1–S7) is provided as a separate file.