GEOMETRIC FEATURE LEARNING FOR FUNCTIONAL DATA VALUED ON THE SYMMETRIC POSITIVE DEFINITE MANIFOLD

Anonymous authors
Paper under double-blind review

ABSTRACT

We here develop a functional neural network, termed `MatFAE`, for learning trajectories on the Riemannian manifold of symmetric positive definite (SPD) matrices. `MatFAE` features intrinsic layers that map manifold-valued functions to Euclidean vector-valued functions, followed by a functional layer that projects them into a finite-dimensional Euclidean space. Unlike most neural networks for discrete-time sequences, `MatFAE` treats each sequence as a continuous function and can therefore encode trajectory dynamics (e.g., first-order derivatives) in its latent representations. Additionally, the morphology of the functional weights in the functional layer offers interpretability by revealing the regions of the input functional data that contribute most to the latent representations. We justify the design principles and properties of each intrinsic layer and detail how matrix factorization is handled during backpropagation. We apply `MatFAE` to a range of fMRI datasets, demonstrating its ability to efficiently learn informative representations from high-dimensional SPD trajectories and its practical value for real-world neuroimaging analysis.

1 INTRODUCTION

Functional data taking values in the Riemannian manifold of symmetric positive-definite (SPD) matrices are increasingly common in scientific and engineering applications. In neuroimaging, time-varying functional connectivity from fMRI is typically represented as trajectories of SPD covariance or correlation matrices (Lurie et al., 2020). In DT-MRI, each voxel is associated with a 3×3 SPD diffusion tensor that, when sampled along white matter tracts, yields SPD-valued functions over a one-dimensional spatial domain (Yuan et al., 2013). In subsurface hydrology and geomechanics, anisotropic hydraulic conductivity (or permeability) is commonly modelled by a 3×3 SPD tensor describing directional flow of groundwater or heat; profiles measured along boreholes or evolving in time during injection, pumping, or thermal loading likewise form SPD-valued functional data (Sanchez-Vila et al., 2006). Other application fields include dynamic texture/action recognition, where each frame or temporal segment is encoded by an SPD descriptor Guo et al. (2013), and structural health monitoring, where covariance matrices of vibration or strain signals from sensor networks are tracked over time.

A natural representation-learning strategy for SPD-matrix trajectories is to interpret each matrix as a graph and model each trajectory as a sequence of graphs, to which one can apply temporal attention mechanisms or recurrent architectures to capture dependencies over time (Chakraborty et al., 2018; Kim et al., 2021). Such sequence-based approaches, however, treat the functional data as discrete-time snapshots and may fail to capture structure in the underlying continuous-time dynamics. We hypothesise that first-order manifold behaviour, which captures how trajectories evolve in direction and speed, encodes additional discriminative and mechanistic information. A functional data perspective on SPD-valued trajectories provides a principled way to capture both the trajectories and their derivatives on the manifold, and to embed this richer geometric and dynamical information into latent representations for downstream tasks such as clustering. A detailed review of related work is given in Appendix A.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Leveraging recent advances in Riemannian neural networks (Huang and Gool, 2017; Chakraborty et al., 2022; Brooks et al., 2019) and functional data analysis (Akeweje and Zhang, 2024; Singh et al., 2025), we introduce `MatFAE`, a neural architecture for representation learning with functional data taking values on the manifold of SPD matrices. To the best of our knowledge, `MatFAE` is the first neural network explicitly designed for this setting. Its geometric layers are intrinsically Riemannian, respecting the non-Euclidean structure of SPD-valued trajectories and inducing a principled geometric bias that yields more efficient and structured representations.

The `MatFAE` architecture is modular. In Section 2, we explain the dimension-reduction module, which maps high-dimensional SPD-valued functional data to lower-dimensional Euclidean functional data. In Section 3, we present the autoencoder module, which learns latent representations of the vector-valued Euclidean functional data and reconstructs the original SPD-valued functional data. Section 4 outlines the computational challenges in training the network, and Section 5 reports extensive numerical studies. All proofs are collected in the appendix.

2 Mapping manifold-valued functions to Euclidean functions

Let $\mathcal{S}_+^m$ denote the manifold of $m \times m$ SPD matrices: $\mathcal{S}_+^m = \{\mathbf{X} \in \mathbb{R}^{m\times m} : \mathbf{X} = \mathbf{X}^\top, \mathbf{X} \succ 0\}$. Given a matrix-valued trajectory $\mathbf{X}(t) \in \mathcal{S}_+^m$ for $t \in [t_0, t_1]$, we below develop a geometric network $\mathcal{F}$ that maps $\mathbf{X}(t)$ to a p-dimensional Euclidean function $\boldsymbol{y}(t) = \mathcal{F}(\mathbf{X}(t)) \in \mathbb{R}^p$. We employ the Log-Euclidean (LE) metric (Arsigny et al., 2007) and the Log-Cholesky (LC) metric (Lin, 2019), both having the bounded-determinant property: for a collection of SPD matrices, the determinant of the Fréchet mean under the metric lies within the range of the input determinants. This property is crucial for the average-pooling layers in our network. Although the affine-invariant metric (Pennec et al., 2006) also satisfies the bounded-determinant property, there is no general closed-form expression for the Fréchet mean, making its use substantially more costly and complicating backpropagation. To avoid confusion, we write $\exp(\mathbf{U})$ for the matrix exponential and $\log(\mathbf{U})$ for the principal matrix logarithm. By contrast, $\mathrm{Exp}_{\mathbf{X}}(\mathbf{U})$ denotes the Riemannian exponential at $\mathbf{X}$ and $\mathrm{Log}_{\mathbf{X}}(\mathbf{Z})$ denotes the Riemannian logarithm at $\mathbf{X}$.

All trainable weights in $\mathcal{F}$ are time-invariant; thus, when no ambiguity arises, we expound the architecture at a fixed t and write $\mathbf{X}$ for $\mathbf{X}(t)$. The manifold $\mathcal{S}_+^m$ is an open cone in $\mathrm{Sym}(m) = \{\mathbf{U} \in \mathbb{R}^{m\times m} : \mathbf{U} = \mathbf{U}^\top\}$, and hence the tangent space $\mathcal{T}_{\mathbf{X}}\mathcal{S}_+^m$ at any point $\mathbf{X} \in \mathcal{S}_+^m$ identifies with $\mathrm{Sym}(m)$: $\mathcal{T}_{\mathbf{X}}\mathcal{S}_+^m = \mathrm{Sym}(m)$, independently of $\mathbf{X}$. Let $\mathfrak{L}(\mathbf{X})$ be the lower triangular matrix of the Cholesky decomposition of $\mathbf{X}$: $\mathbf{X} = \mathfrak{L}(\mathbf{X})\mathfrak{L}(\mathbf{X})^\top$, and $\lfloor\mathfrak{L}(\mathbf{X})\rfloor$ the strictly lower triangular matrix of $\mathfrak{L}(\mathbf{X})$. Define the log-Cholesky coordinate map $\Phi : \mathcal{S}_+^m \to \mathbb{R}^{m(m+1)/2}$, $\Phi(\mathbf{X}) = (\mathrm{vnz}(\mathfrak{L}(\mathbf{X})), \log(\mathrm{vd}(\mathfrak{L}(\mathbf{X}))))^\top$, where $\mathrm{vnz}(\mathfrak{L}(\mathbf{X}))$ is the vector of the non-zero entries of $\lfloor\mathfrak{L}(\mathbf{X})\rfloor$, and $\mathrm{vd}(\mathfrak{L}(\mathbf{X}))$ is the vector of the diagonal entries of $\mathfrak{L}(\mathbf{X})$. The LC metric makes Φ a global Riemannian isometry: equipping $\Phi(\mathbf{X})$ with the standard Euclidean inner product and pulling it back by Φ defines the LC metric:

$$\langle \mathbf{U}, \mathbf{V}\rangle_{\mathbf{X}}^{\mathrm{LC}} = \langle (D_{\mathbf{X}}\Phi)(\mathbf{U}),\ (D_{\mathbf{X}}\Phi)(\mathbf{V})\rangle_2, \quad \mathbf{U}, \mathbf{V} \in \mathcal{T}_{\mathbf{X}}\mathcal{S}_+^m = \mathrm{Sym}(m),$$

where $(D_{\mathbf{X}}\Phi)(\mathbf{U}) = \frac{d}{ds}|_{s=0}\Phi(\mathbf{X} + s\mathbf{U})$ is the Fréchet derivative of Φ at the point $\mathbf{X}$ to the tangent direction $\mathbf{U}$. Likewise, the LE metric is the pullback of the Frobenius metric on $\mathrm{Sym}(m)$ via the principal matrix logarithm $\log : \mathcal{S}_+^m \to \mathrm{Sym}(m)$; for $\mathbf{X} \in \mathcal{S}_+^m$ and $\mathbf{U}, \mathbf{V} \in \mathcal{T}_{\mathbf{X}}\mathcal{S}_+^m$, the LE inner product at $\mathbf{X}$ is

$$\langle \mathbf{U}, \mathbf{V}\rangle_{\mathbf{X}}^{\mathrm{LE}} = \langle (D_{\mathbf{X}}\log)(\mathbf{U}),\ (D_{\mathbf{X}}\log)(\mathbf{V})\rangle_F,$$

where $(D_{\mathbf{X}}\log)(\mathbf{U}) = \frac{d}{ds}|_{s=0}\log(\mathbf{X} + s\mathbf{U})$ is the Fréchet derivative of $\log$ at the point $\mathbf{X}$ to the tangent direction $\mathbf{U}$. Both Riemannian manifolds $(\mathcal{S}_+^m, \langle\cdot,\cdot\rangle^{\mathrm{LC}})$ and $(\mathcal{S}_+^m, \langle\cdot,\cdot\rangle^{\mathrm{LE}})$ are flat (zero sectional curvature), geodesically complete, and simply connected; hence each is a Cartan-Hadamard manifold. Further properties of the LC and LE metrics are provided in Appendix B.

According to Theorem 1 in Singh et al. (2025), on Cartan-Hadamard manifolds we can linearize the manifold-valued functions by mapping them into Euclidean space via the (globally defined) Riemannian logarithm map. The linearized functions can then be processed by the FAEclust functional network, which guarantees universal approximation. However, a direct mapping of $\mathbf{X}(t)$ onto $\mathrm{Sym}(m)$ produces $\boldsymbol{y}(t)$ with dimension $m(m+1)/2$, which imposes a heavy training cost for FAEclust. Therefore, we here develop a modular geometric network that compresses the trajectory

$\mathbf{X}(t)$ to a lower-order SPD trajectory $\mathbf{X}_4(t) \in \mathcal{S}_+^{m_2}$ with $m_2 < m$, and then applies the Riemannian logarithm to map the trajectory $\mathbf{X}_4(t)$ to the Euclidean space $\mathrm{Sym}(m_2)$ for subsequent processing by any functional neural network. This dimensionality reduction substantially decreases computational burden while preserving the relevant geometry.

An overview of the network architecture is provided in Figure 1. The two BiMap layers perform dimensionality reduction, the Pooling layer aggregates SPD outputs across heads, and the Activation layer appropriately regularize the SPD matrix. The Logarithm layer performs Riemannian logarithm, mapping each head $\mathbf{X}_4^h(t)$ to $\mathbf{Y}^h(t)$: $\mathbf{Y}^h(t) = \mathrm{Log}_{\mathbf{I}}^{\mathrm{LC}}(\mathbf{X}_4^h(t))$ or $\mathbf{Y}^h(t) = \mathrm{Log}_{\mathbf{I}}^{\mathrm{LE}}(\mathbf{X}_4^h(t))$, where $\mathbf{I}$ is the identity matrix. The explicit expressions for $\mathrm{Log}_{\mathbf{I}}^{\mathrm{LC}}(\cdot)$ and $\mathrm{Log}_{\mathbf{I}}^{\mathrm{LE}}(\cdot)$ are provided in Eqs. (6) and (10), respectively. In the Concat layer, we vectorize the lower-triangular entries of $\mathbf{Y}^h(t)$ and concatenate them across heads to obtain a vector-valued Euclidean function $\boldsymbol{y}(t)$. We now provide a rigorous justification for the BiMap, Pooling, and Activation layers and detail their properties.

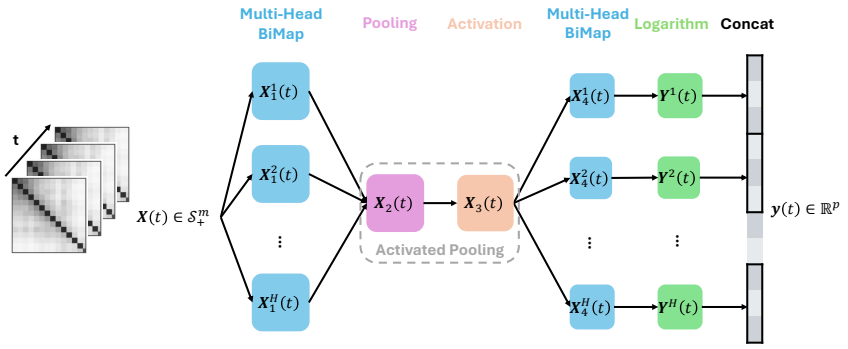

Figure 1: The two BiMap layers reduce the matrix dimension from m to $m_1 (< m)$, and further from m_1 to $m_2 (< m_1)$; that is, $\mathbf{X}_1^h(t) \in \mathcal{S}_+^{m_1}$ and $\mathbf{X}_4^h(t) \in \mathcal{S}_+^{m_2}$ for $1 \le h \le H$. The Pooling, Activation and Logarithm layers are geometric layers that depend on the Riemannian metric on $\mathcal{S}_+^m$. The Pooling layer performs manifold-average pooling of the SPD matrices $\{\mathbf{X}_1^h(t) : 1 \le h \le H\}$; the Activation layer applies an activation on $\mathbf{X}_2(t)$, and the Logarithm layer maps each $\mathbf{X}_4^h(t)$ to the tangent space via the Riemannian logarithmic map $\mathrm{Log}_{\mathbf{I}}^{\mathrm{LC}}(\cdot)$ or $\mathrm{Log}_{\mathbf{I}}^{\mathrm{LE}}(\cdot)$. Finally, we vectorize the lower-triangular entries of $\mathbf{Y}^h(t)$ and concatenate them across heads to obtain a vector-valued function $\boldsymbol{y}(t)$.

2.1 Multi-head BiMap

Huang and Gool (2017) introduced SPDNet, a deep architecture that operates directly on SPD matrices and preserves positive-definiteness throughout. Its core linear block (i.e., the BiMap layer) applies a congruence transformation: $\mathbf{X}_k = \mathbf{W}_k \mathbf{X}_{k-1} \mathbf{W}_k^\top$, where $\mathbf{X}_{k-1} \in \mathcal{S}_+^{d_{k-1}}$, and $\mathbf{W}_k \in \mathbb{R}^{d_k \times d_{k-1}}$ $(d_k < d_{k-1})$ is a full row rank weight matrix with orthonormal rows. The bilinear congruence ensures that the output remains SPD, i.e., $\mathbf{X}_k \in \mathcal{S}_+^{d_k}$, while reducing dimension from d_{k-1} to d_k. Stacking multiple BiMap layers yields progressively more compact and task-discriminative SPD representations.

For the $\mathcal{S}_+^m$-valued function $\mathbf{X}(t)$, a natural extension of the bilinear mapping is $\mathbf{X}_1(t) = \mathbf{W}_1 \mathbf{X}(t) \mathbf{W}_1^\top$. However, time-varying SPD signals almost never commute across time (i.e., $\mathbf{X}(t_1)\mathbf{X}(t_2) \neq \mathbf{X}(t_2)\mathbf{X}(t_1)$), and therefore no single global projection $\mathbf{W}_1$ can align the covariance geometry uniformly over t. When regimes switch and eigenvalues cross, causing the spectral subspaces to drift, any fixed $\mathbf{W}_1$ will inevitably overfit some intervals and underfit others. We therefore introduce a multi-head BiMap layer with H parallel bilinear maps: $\mathbf{X}_1^h(t) = \mathbf{W}_1^h \mathbf{X}(t) (\mathbf{W}_1^h)^\top$, $h = 1, \ldots, H$, each with its own full row rank $\mathbf{W}_1^h$.

When $\mathbf{X}(t)$ for different t values do not share a common eigenbasis, different heads approximate different (temporally) local joint structures, improving conditioning and reducing bias compared with a single global projection. Different heads naturally specialize to complementary, time-dependent re-

lationships (e.g., distinct temporal regimes, frequency bands, or sensor groups). Their SPD outputs are then fused by an SPD-preserving pooling operator in the Pooling layer. This design also mirrors 1-D CNNs along the temporal axis: each $\mathbf{W}_1^h$ serves as a shared temporal filter, and multiple heads enhance expressive power by capturing diverse temporal patterns while retaining weight sharing.

2.2 POOLING

The Pooling layer aggregates the SPD inputs $\{\mathbf{X}_1^h : 1 \le h \le H\}$. A natural choice is the Fréchet mean on $\mathcal{S}_+^{m_1}$ under either the LC or LE geometry:

$$\mathbf{X}_2 = \underset{\mathbf{X}\in\mathcal{S}_+^{m_1}}{\arg\min} \sum_{h=1}^{H} d_g(\mathbf{X}, \mathbf{X}_1^h)^2,$$

where d_g is the geodesic distance. Under either the LC metric or the LE metric, the Fréchet mean has a closed form (Eqs. (7) and (11)). Conceptually, this operation is the manifold analogue of global average pooling: it compresses H inputs into one SPD representative while strictly preserving positive-definiteness. A caveat is that the congruence mapping $\mathbf{W}_1^h\mathbf{X}(\mathbf{W}_1^h)^\top$ does not commute with the matrix logarithm/exponential in the LE metric: in general,

$$\log(\mathbf{W}_1^h\mathbf{X}(\mathbf{W}_1^h)^\top) \neq \mathbf{W}_1^h \log(\mathbf{X})(\mathbf{W}_1^h)^\top \quad \text{and} \quad \exp(\mathbf{W}_1^h\mathbf{X}(\mathbf{W}_1^h)^\top) \neq \mathbf{W}_1^h \exp(\mathbf{X})(\mathbf{W}_1^h)^\top.$$

Likewise, for the LC metric, the Cholesky factor $\mathfrak{L}(\mathbf{W}_1^h\mathbf{X}(\mathbf{W}_1^h)^\top)$ depends nonlinearly on both $\mathbf{W}_1^h$ and $\mathbf{X}$. Consequently, the Fréchet mean computed in the Pooling layer is neither invariant nor equivariant to dimension-reducing congruence actions applied to $\mathbf{X}(t)$. We detail this lack of invariance/equivariance below.

A canonical family of symmetries of $\mathcal{S}_+^m$ is given by the congruence action of $\mathrm{GL}(m)$: any invertible matrix $\mathbf{U} \in \mathrm{GL}(m)$ defines the symmetry $\mathbf{X} \mapsto \mathbf{U}\mathbf{X}\mathbf{U}^\top$, acting on all points $\mathbf{X} \in \mathcal{S}_+^m$ (López et al., 2021). (1) When $\mathbf{U}$ is an element of $\mathcal{S}_+^m$, the symmetry $\mathbf{X} \mapsto \mathbf{U}\mathbf{X}\mathbf{U}^\top$ is a generalization of the Euclidean *translation*, fixing no points of $\mathcal{S}_+^m$. (2) When $\mathbf{U} \in \mathrm{O}(m)$ is an orthogonal matrix, the symmetry $\mathbf{X} \mapsto \mathbf{U}\mathbf{X}\mathbf{U}^\top$ is conjugation by $\mathbf{U}$, and hence fixes the base point $\mathbf{I} = \mathbf{U}\mathbf{I}\mathbf{U}^\top = \mathbf{U}\mathbf{U}^\top = \mathbf{I}$. The stabilizer of $\mathbf{I}$ is precisely $\mathrm{O}(m)$: elements with $\det(\mathbf{U}) = 1$ are *$\mathcal{S}_+^m$-rotations*, and elements with $\det(\mathbf{U}) = -1$ are *$\mathcal{S}_+^m$-reflections*.

Proposition 1. *The LE Fréchet mean $\mathbf{X}_2$ is invariant or equivariant under orthogonal congruence only under pathological conditions. The LC Fréchet mean is neither invariant nor equivariant under orthogonal congruence. Moreover, under the LC and LE geometries, the Fréchet mean $\mathbf{X}_2$ is generally neither invariant nor equivariant under actions of the isometry group on $\mathcal{S}_+^m$.*

Max pooling on $\mathcal{S}_+^{m_1}$ Unlike Euclidean space, the manifold $\mathcal{S}_+^{m_1}$ admits neither canonical global coordinates nor a natural total order. Therefore, a "max" operation is not intrinsically defined: any elementwise maximum defined in a particular chart (e.g., $\log$ or Φ) is chart-dependent and generally fails to be equivariant under manifold isometries. An intrinsic alternative is to define a scalar score $s : \mathcal{S}_+^{m_1} \to \mathbb{R}$ and select the head $\mathbf{X}_1^h$ with the highest score $s(\mathbf{X}_1^h)$. However, because features pass through the congruence map $\mathbf{X} \mapsto \mathbf{W}_1^h\mathbf{X}(\mathbf{W}_1^h)^\top$, no non-trivial choice of s yields invariance/equivariance to general matrix congruence; this follows from the same obstruction discussed in Appendix C. Equivariance is attainable only in pathological cases (e.g., orthogonal conjugation with spectral scores, or specific metric-based selections with matched transformations of any reference). Moreover, the $\max$ operation is non-smooth and therefore less amenable to optimization. For these reasons, `MatFAE` adopts average (Fréchet) pooling.

Although the BiMap's rectangular congruence deprives the Pooling layer of invariance or equivariance guarantees, it remains a highly efficient dimension-reduction operator: it preserves SPD by construction and needs only two matrix-matrix products, completely avoiding eigendecomposition.

2.3 ACTIVATION

Conventional deep architectures interleave pointwise nonlinearities $a : \mathbb{R} \to \mathbb{R}$ between linear layers (e.g., ReLU). Such activations are non-linear and, in most cases, non-expansive with respect to the Euclidean norm (i.e., $\|a(\boldsymbol{x}) - a(\boldsymbol{y})\|_2 \le L\|\boldsymbol{x} - \boldsymbol{y}\|_2$ with $L \le 1$); certain choices (e.g.,

the logistic sigmoid) are strictly contractive. Non-expansiveness helps control the network's overall Lipschitz constant, improving numerical stability, robustness to perturbations, and generalization by biasing toward smoother mappings. The non-linearities between layers prevent the deep neural networks from collapsing to a single fully connected layer. Below we prove that the Fréchet-mean pooling operator is, in general, neither contractive nor non-expansive (Proposition 2 and Proposition 3, which are different from Theorem 1 and Proposition 2 in Chakraborty et al. (2022)). However, the Fréchet-mean pooling operator itself is non-linear (Proposition 4). Therefore, a repeated block "BiMap $\to$ Pooling $\to$ BiMap $\to$ Pooling" will not collapse to a single "BiMap $\to$ Pooling" block.

Proposition 2. *Fix any spectral band $0 < \alpha \leq \beta < \infty$ and consider the subset $\mathcal{X}_{\alpha,\beta} = \{\mathbf{X} \in \mathcal{S}_+^m : \alpha\mathbf{I} \preceq \mathbf{X} \preceq \beta\mathbf{I}\}$. Under the LE metric, the mapping $F : \mathcal{S}_+^m \mapsto \mathcal{S}_+^{m_1}$,*

$$F(\mathbf{X}; \{\mathbf{W}_1^h\}_{h=1}^H) = \mathbb{E}_{LE}(\mathbf{W}_1^1\mathbf{X}(\mathbf{W}_1^1)^\top, \ldots, \mathbf{W}_1^H\mathbf{X}(\mathbf{W}_1^H)^\top) = \exp(\frac{1}{H}\sum_{h=1}^H \log(\mathbf{W}_1^h\mathbf{X}(\mathbf{W}_1^h)^\top),$$

is Lipschitz on $\mathcal{X}_{\alpha,\beta}$ with constant $L \leq \beta/\alpha$. In particular, F is neither contractive nor non-expansive.

Proposition 3. *Fix any spectral band $0 < \alpha \leq \beta < \infty$ and consider the subset $\mathcal{X}_{\alpha,\beta} = \{\mathbf{X} \in \mathcal{S}_+^m : \alpha\mathbf{I} \preceq \mathbf{X} \preceq \beta\mathbf{I}\}$. Under the LC metric, the mapping $F : \mathcal{S}_+^m \mapsto \mathcal{S}_+^{m_1}$,*

$$F(\mathbf{X}; \{\mathbf{W}_1^h\}_{h=1}^H) = \mathbb{E}_{LC}(\mathbf{W}_1^1\mathbf{X}(\mathbf{W}_1^1)^\top, \ldots, \mathbf{W}_1^H\mathbf{X}(\mathbf{W}_1^H)^\top) = \Phi^{-1}(\frac{1}{H}\sum_{h=1}^H \Phi(\mathbf{W}_1^h\mathbf{X}(\mathbf{W}_1^h)^\top),$$

is Lipschitz with constant $L \leq \frac{2\beta^{3/2}}{\alpha}\max\{1, \alpha^{-1}\}$. In particular, F is neither contractive nor non-expansive.

Proposition 4. *The pooling map is itself nonlinear. Without any non-linear activation after the Pooling layer, the repeated block "BiMap $\to$ Pooling $\to$ BiMap $\to$ Pooling" will not collapse to a single "BiMap $\to$ Pooling" block.*

Proposition 4 implies that, if contractivity is not required, the Activation layer can be omitted. The architecture in Figure 1 then reduces to "BiMap $\to$ Pooling $\to$ BiMap $\to$ Logarithm $\to$ Concat". However, the Fréchet-mean pooling is only Lipschitz, neither contractive nor non-expansive (Propositions 2 and 3). Therefore, when stability or spectral control is desired, inserting a contractive activation after pooling will enforce a uniform Lipschitz bound, curb eigenvalue inflation, and improve numerical stability.

The ReEig layer introduced in Huang and Gool (2017) performs a spectral rectification by eigendecomposing $\mathbf{X}_2 = \mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^\top$, replacing each eigenvalue λ_i with $\max\{\lambda_i, \epsilon\}$, and reconstructing $\mathbf{X}_3 = \mathbf{Q}\max\{\mathbf{\Lambda}, \epsilon\mathbf{I}\}\mathbf{Q}^\top$. However, the ReEig activation operator is not scale-invariant (positively homogeneous); for $c > 0$, we will have $c\mathbf{X}_3 \neq \mathbf{Q}\max\{c\mathbf{\Lambda}, \epsilon\mathbf{I}\}\mathbf{Q}^\top$, whenever a scaled eigenvalue crosses the threshold. We can modify the ReEig operation to be scale-invariant:

$$\mathbf{X}_3 = s(\mathbf{X}_2)\mathbf{Q}\exp(a(\log(\mathbf{\Lambda}) - \log(s(\mathbf{X}_2))\mathbf{I}))\mathbf{Q}^\top,$$

where $a(\cdot)$ is elementwise activation, and $s(\mathbf{X}_2)$ is any 1-homogeneous scale functional, e.g., $s(\mathbf{X}_2) = \frac{1}{m_1}\text{trace}(\mathbf{X}_2)$. The map is positively homogeneous: $\log(c\mathbf{X}_2) = \mathbf{Q}[\log(\mathbf{\Lambda}) + \log(c)\mathbf{I}]\mathbf{Q}^\top$ and $s(c\mathbf{X}_2) = c \times s(\mathbf{X}_2)$. However, embedding $s(\mathbf{X}_2)$ inside the spectral nonlinearity couples the scale and eigenvalue paths in backpropagation, requiring extra spectral Fréchet operations and raising the computational burden ($\approx \mathcal{O}(m_1^3)$). Another type of activation, introduced by Zhang et al. (2018), directly applies the scalar function $\exp(\cdot)$, $\sinh(\cdot)$ or $\cosh(\cdot)$ elementwise to $\mathbf{X}_2$. However, $\exp(\cdot)$ and $\cosh(\cdot)$ grow rapidly; their derivatives can trigger exploding activations/gradients and inflate condition numbers. Moreover, elementwise maps provide no direct spectral control and can mix scales, obscuring links to the underlying process.

Motivated by the need to control the Lipschitz bound and by the fact that our pooling operator is already non-linear, we develop an intrinsic geodesic-shrinkage activation on $\mathcal{S}_+^{m_1}$, which preserves scale behavior while yielding tractable derivatives.

Geodesic shrinkage Let ϕ denote either the principal matrix logarithm $\log$ or the log-Cholesky coordinate map Φ. The Pooling layer computes the Fréchet mean in ϕ-coordinates via

$$F(\mathbf{X}; \{\mathbf{W}_1^h\}_{h=1}^H) = \phi^{-1}(\frac{1}{H}\sum_{h=1}^H \phi(\mathbf{W}_1^h\mathbf{X}(\mathbf{W}_1^h)^\top)).$$

We want to regularize the Fréchet mean $\mathbf{X}_2 = F(\mathbf{X}; \{\mathbf{W}_1^h\}_{h=1}^H)$ by moving $\mathbf{X}_2$ back toward the identity matrix $\mathbf{I}$ along the unique geodesic joining $\mathbf{I}$ to $\mathbf{X}_2$, stopping at a proportion $0 < \alpha < 1$ of the total geodesic length. Because ϕ maps $\mathbf{I}$ to the origin of the coordinate space: $\phi(\mathbf{I}) = \mathbf{0}$, geodesic distances from $\mathbf{I}$ reduce to Euclidean norms: $d_{\text{LE}}(\mathbf{I}, \mathbf{X}_2) = \|\log(\mathbf{X}_2)\|_F$ under the LE metric, and $d_{\text{LC}}(\mathbf{I}, \mathbf{X}_2) = \|\Phi(\mathbf{X}_2)\|_2$ under the LC metric. Define the activation output $\mathbf{X}_3$ as the point at distance $\alpha d_g(\mathbf{I}, \mathbf{X}_2)$ from $\mathbf{I}$ along that geodesic. Since geodesics are straight lines in ϕ-coordinates, this construction has a particularly simple coordinate description:

$$\phi(\mathbf{X}_3) = (1-\alpha)\phi(\mathbf{I}) + \alpha\phi(\mathbf{X}_2) = \alpha\phi(\mathbf{X}_2),$$

or, equivalently, $\mathbf{X}_3 = \phi^{-1}(\alpha\phi(\mathbf{X}_2))$. In words: we shrink the coordinate vector toward zero (the image of $\mathbf{I}$) by a factor α, and then map back to the manifold. This yields $d_g(\mathbf{I}, \mathbf{X}_3) = \alpha d_g(\mathbf{I}, \mathbf{X}_2)$. Hence the operation contracts distances from the identity exactly by α. The effect is directly analogous to weight decay in Euclidean networks, but carried out in intrinsic coordinates of the manifold.

To keep overhead negligible, rather than "pool, then activate," we combine the Pooling layer and the Activation layer into one layer, namely, the "Activated Pooling" layer in Figure 1. The activated-pooling computes the average in the ϕ-chart, shrinks the average by α, and maps the shrinked average back to the manifold: $\mathbf{X}_3 = \phi^{-1}(\frac{\alpha}{H}\sum_{h=1}^H \phi(\mathbf{W}_1^h \mathbf{X} (\mathbf{W}_1^h)^\top))$. This eliminates the intermediate $\mathbf{X}_2$. The only extra computation relative to plain pooling is a scalar multiplication by α in the ϕ-domain.

2.4 Concatenation

Substituting the intermediate layers, the composite mapping from the input $\mathbf{X}(t)$ to $\mathbf{Y}^h(t)$ is:

$$\mathbf{Y}^h(t) = \text{Log}_{\mathbf{I}}^g \left(\mathbf{W}_4^h \phi^{-1}\left(\frac{\alpha}{H}\sum_{k=1}^H \phi\left(\mathbf{W}_1^k \mathbf{X}(t)(\mathbf{W}_1^k)^\top\right)\right)(\mathbf{W}_4^h)^\top \right). \tag{1}$$

For each head $h = 1, \ldots, H$ and time t, the Concatenation layer maps a symmetric matrix $\mathbf{Y}^h(t) \in \text{Sym}(m_2)$ to a vector $\text{vl}(\mathbf{Y}^h(t)) \in \mathbb{R}^{m_2(m_2+1)/2}$, and then concatenates the vectors across heads. We equip $\text{Sym}(m_2)$ with the Frobenius inner product and $\mathbb{R}^{m_2(m_2+1)/2}$ with the Euclidean inner product and require $\text{vl}(\cdot)$ to be an isometry: $\langle \mathbf{Y}_1, \mathbf{Y}_2 \rangle_F = \langle \text{vl}(\mathbf{Y}_1), \text{vl}(\mathbf{Y}_2) \rangle_2$, for all $\mathbf{Y}_1, \mathbf{Y}_2$. Isometry guarantees norm preservation and, crucially for backpropagation, it makes the adjoint equal the inverse. Fix any indexing $\kappa : \{(i,j) : 1 \le j \le i \le m_2\} \mapsto \{1, \ldots, m_2(m_2+1)/2\}$ (e.g., lexicographic). Define $[\text{vl}(\mathbf{Y})]_{\kappa(i,i)} = Y_{i,i}, \quad [\text{vl}(\mathbf{Y})]_{\kappa(i,j)} = \sqrt{2} Y_{i,j}$, for $i > j$. It is easy to prove that $\text{vl}(\cdot)$ is an isometry, and the adjoint equals the inverse. Finally, the output of the network is $\boldsymbol{y}(t)^\top = [\text{vl}(\mathbf{Y}^1(t))^\top, \ldots, \text{vl}(\mathbf{Y}^H(t))^\top]$.

3 Functional autoencoder with geometry-aware layers

The matrix-to-vector network in Figure 1 is modular: it can be paired with any network that accept vector-valued functions. We here adapt the FAEclust architecture (Singh et al., 2025) to learn a latent embedding of the Euclidean vector-valued function $\boldsymbol{y}(t)$ and, via a geometry-aware decoder, to reconstruct the original SPD trajectory $\mathbf{X}(t)$. FAEclust comprises an encoder $\mathcal{E}$ and a decoder $\mathcal{D}$ (see Figure 2). Let $\mathcal{H}([t_0, t_1]; \mathbb{R}^p)$ denote the separable Hilbert space of square-integrable functions that are defined on $[t_0, t_1]$ and taking values in $\mathbb{R}^p$. Given a p-variate function $\boldsymbol{y} \in \mathcal{H}([t_0, t_1], \mathbb{R}^p)$, the encoder maps the entire trajectory to a latent multivariate data point $\boldsymbol{x} = \mathcal{E}(\boldsymbol{y}) \in \mathbb{R}^s$; the decoder then reconstructs a functional trajectory $\hat{\boldsymbol{y}} = \mathcal{D}(\boldsymbol{x}) \in \mathcal{H}([t_0, t_1], \mathbb{R}^p)$. Cluster analysis is performed on the embedded data in the latent space. Unlike a conventional MLP autoencoder, FAEclust introduces four functional layers: one placed at the encoder entrance and three at the decoder exit, which act on functions rather than vectors. The feedforward equation for the functional layer in the encoder is given by

$$\boldsymbol{x}^{(1)} = a\left(\int_{t_0}^{t_1} \boldsymbol{W}^{(1)}(t)\boldsymbol{y}(t)dt + \boldsymbol{b}\right), \tag{2}$$

where $\boldsymbol{W}^{(1)}(t)$ is a functional weight matrix, $\boldsymbol{b}$ is a bias vector, and a is an activation function. $\boldsymbol{x}^{(1)}$ is then fed into the MLP autoencoder. The equations for the three functional layers in the decoder

are:

$$\hat{\boldsymbol{y}}^{(1)}(t) = a(\boldsymbol{\mathcal{W}}^{(1)}(t)\hat{\boldsymbol{x}} + \boldsymbol{b}_1(t)),$$
$$\hat{\boldsymbol{y}}^{(2)}(t) = a(\boldsymbol{\mathcal{W}}^{(2)}(t)\hat{\boldsymbol{y}}^{(1)}(t) + \boldsymbol{b}_2(t)),$$
$$\hat{\boldsymbol{y}}(t) = \boldsymbol{\mathcal{W}}^{(3)}(t)\hat{\boldsymbol{y}}^{(2)}(t),$$

where $\hat{\boldsymbol{x}}$ is the output of the MLP autoencoder, $\{\boldsymbol{\mathcal{W}}^{(l)}(t), l = 1, 2, 3\}$ are functional weight matrices, and $\{\boldsymbol{b}_1(t), \boldsymbol{b}_2(t)\}$ are functional biases. For the output layer, the activation function is linear, and there is no functional bias.

For a matrix $\mathbf{U} \in \mathrm{Sym}(m)$, let $\mathcal{V}(\mathbf{U}) = (\mathrm{vec}(\lfloor \mathbf{U} \rfloor), \mathrm{vec}(\mathrm{diag}(\mathbf{U}))) \in \mathbb{R}^{m(m+1)/2}$ denote the vector of the lower triangular entries of $\mathbf{U}$. Note that $\mathcal{V}$ is different from the log-Cholesky coordinate map Φ. The inverse $\mathcal{V}^{-1}$ maps a vector $\boldsymbol{u} \in \mathbb{R}^{m(m+1)/2}$ back to an $m \times m$ matrix in $\mathrm{Sym}(m)$: $\mathbf{U} = \mathcal{V}^{-1}(\mathcal{V}(\mathbf{U}))$. We can directly apply FAEclust and train the network to output $\hat{\boldsymbol{y}}(t) \in \mathbb{R}^{m(m+1)/2}$ that approximates $\mathcal{V}(\mathrm{Log}_{\mathbf{I}}^{\mathrm{LC}}(\mathbf{X}(t)))$ or $\mathcal{V}(\mathrm{Log}_{\mathbf{I}}^{\mathrm{LE}}(\mathbf{X}(t)))$. In practice, this is computationally prohibitive: each functional weight matrix $\boldsymbol{\mathcal{W}}^{(l)}(t)$ will have size about $m^2 \times m^2$, yielding $\mathcal{O}(m^4)$ parameters per layer and a heavy training burden. We cut computation by decoding in a low-dimensional tangent space, then lifting via conjugation with thin, column-orthonormal matrices, giving an injective isometry for the Frobenius inner product. Let $\hat{\boldsymbol{y}}^{(1)}(t) \in \mathbb{R}^{p_1(p_1+1)/2}$, where $p_1 < m$, and $\mathcal{V}^{-1}(\hat{\boldsymbol{y}}^{(1)}(t)) \in \mathrm{Sym}(p_1)$ is a trajectory in the tangent space of $\mathcal{S}_+^{p_1}$. Let $\mathbf{M}_1 \in \mathbb{R}^{p_2 \times p_1}$ and $\mathbf{M}_2 \in \mathbb{R}^{m \times p_2}$ be two thin matrices of orthonormal columns. The modified functional layers are

$$\hat{\boldsymbol{y}}^{(1)}(t) = a(\boldsymbol{\mathcal{W}}^{(1)}(t)\hat{\boldsymbol{x}} + \boldsymbol{b}_1(t)),$$
$$\hat{\boldsymbol{y}}^{(2)}(t) = a(\boldsymbol{\mathcal{W}}^{(2)}(t)\mathcal{V}(\mathbf{M}_1\mathcal{V}^{-1}(\hat{\boldsymbol{y}}^{(1)}(t))\mathbf{M}_1^\top) + \boldsymbol{b}_2(t)), \quad (3)$$
$$\hat{\mathbf{X}}(t) = \mathrm{Exp}_{\mathbf{I}}^{g}(\mathbf{M}_2\mathcal{V}^{-1}(\hat{\boldsymbol{y}}^{(2)}(t))\mathbf{M}_2^\top), \quad (4)$$

where $\hat{\mathbf{X}}(t) \in \mathcal{S}_+^m$ is the reconstructed SPD trajectory, and $\mathrm{Exp}_{\mathbf{I}}^{g}$ is the Riemannian exponential at $\mathbf{I}$ under the chosen metric g (i.e., LC or LE). In Eq. 3, we lift from $\mathrm{Sym}(p_1)$ to $\mathrm{Sym}(p_2)$ by conjugation before applying the nonlinearity in vectorized form, and in Eq. 4 we lift again to $\mathrm{Sym}(m)$, and map back to the manifold. In line with the output layer of FAEclust, Eq. 4 only involves a bilinear mapping (in the tangent space).

4 Network Training

For a Riemannian metric g on $\mathcal{S}_+^m$ (i.e., the LC or LE metric), define $\mathcal{H}_g([t_0, t_1]; \mathcal{S}_+^m) = \{\mathbf{X} : [t_0, t_1] \to \mathcal{S}_+^m \text{ measurable} \mid \int_{t_0}^{t_1} d_g(\mathbf{X}(t), \mathbf{I})^2 dt < \infty\}$. The matrix-to-vector module is the mapping $\mathcal{F}$: $\mathbf{X} \in \mathcal{H}_g([t_0, t_1]; \mathcal{S}_+^m) \mapsto \boldsymbol{y} = \mathcal{F}(\mathbf{X}) \in \mathcal{H}([t_0, t_1]; \mathbb{R}^p)$, and the `MatFAE` network is the mapping $\mathcal{D} \circ \mathcal{E} \circ \mathcal{F}$: $\mathbf{X} \in \mathcal{H}_g([t_0, t_1]; \mathcal{S}_+^m) \mapsto \hat{\mathbf{X}} = \mathcal{D}(\mathcal{E}(\mathcal{F}(\mathbf{X}))) \in \mathcal{H}_g([t_0, t_1]; \mathcal{S}_+^m)$. Given a set of $\mathcal{S}_+^m$-valued trajectories $\{\mathbf{X}_i \in \mathcal{H}_g([t_0, t_1]; \mathcal{S}_+^m)\}_{i=1}^n$, `MatFAE` produces reconstructions $\{\hat{\mathbf{X}}_i \in \mathcal{H}_g([t_0, t_1]; \mathcal{S}_+^m)\}_{i=1}^n$. Under the LE metric, the reconstruction loss is

$$\mathcal{L}_r = \sum_{i=1}^{n} \int_{t_0}^{t_1} d_{\mathrm{LE}}(\mathbf{X}_i(t), \hat{\mathbf{X}}_i(t))^2 dt = \sum_{i=1}^{n} \int_{t_0}^{t_1} \| \log(\mathbf{X}_i(t)) - \mathbf{M}_2\mathcal{V}^{-1}(\hat{\boldsymbol{y}}_i^{(2)}(t))\mathbf{M}_2^\top \|_F^2 dt.$$

Under the LC metric, the reconstruction loss $\mathcal{L}_r$ is

$$\sum_{i=1}^{n} \int_{t_0}^{t_1} d_{\mathrm{LC}}(\mathbf{X}_i(t), \hat{\mathbf{X}}_i(t))^2 dt = \sum_{i=1}^{n} \int_{t_0}^{t_1} \| \lfloor \mathfrak{L}(\mathbf{X}_i(t)) \rfloor - \lfloor \mathbf{M}_2\mathcal{V}^{-1}(\hat{\boldsymbol{y}}_i^{(2)}(t))\mathbf{M}_2^\top \rfloor \|_F^2 dt$$
$$+ \sum_{i=1}^{n} \int_{t_0}^{t_1} \| \log(\mathrm{diag}(\mathfrak{L}(\mathbf{X}_i(t)))) - \frac{1}{2}\mathrm{diag}(\mathbf{M}_2\mathcal{V}^{-1}(\hat{\boldsymbol{y}}_i^{(2)}(t))\mathbf{M}_2^\top) \|_F^2 dt.$$

The training objective function of `MatFAE` comprises three components: (1) $\mathcal{L}_r$, the reconstruction loss; (2) $\mathcal{L}_w$, a regularization term on the functional weights and functional biases $\{\boldsymbol{\mathcal{W}}^{(1)}(t), \boldsymbol{\mathcal{W}}^{(1)}(t), \boldsymbol{\mathcal{W}}^{(2)}(t), \boldsymbol{\mathcal{W}}^{(3)}(t), \boldsymbol{b}_1(t), \boldsymbol{b}_2(t)\}$; and (3) $\mathcal{L}_{\mathrm{div}}$, a row-space decorrelation penalty on the congruence weight matrices $\{\mathbf{W}_1^h\}_{h=1}^H$ and $\{\mathbf{W}_4^h\}_{h=1}^H$, The detailed formulation of $\mathcal{L}_w$ is given in (Singh et al., 2025), while the formulation of $\mathcal{L}_{\mathrm{div}}$ is provided in Appendix F.

Riemannian gradient `MatFAE` involves a few matrix parameters that are constrained to Stiefel manifolds. We here briefly outline the gradient-descent procedure for optimizing them. Let $\mathrm{St}(m,p) = \{\mathbf{M} \in \mathbb{R}^{m\times p} : \mathbf{M}^\top\mathbf{M} = \mathbf{I}\}$ denote the Stiefel manifold of $m \times p$ matrices with orthonormal columns (i.e., $m > p$). For any smooth mapping $f : \mathrm{St}(m,p) \to \mathbb{R}$, let $\mathbf{D} = \nabla_{\mathbf{M}} f(\mathbf{M})$ denote the normal Euclidean gradient. Then project $\mathbf{D}$ onto the tangent space to get the Riemannian gradient: $\Delta = \mathbf{D} - \frac{1}{2}\mathbf{M}(\mathbf{M}^\top\mathbf{D} + \mathbf{D}^\top\mathbf{M})$. For a given step size $\eta > 0$, compute a thin QR factorization of $\mathbf{M} + \eta(-\Delta)$: $\mathbf{M} - \eta\Delta = \mathbf{QR}$; then the updated $\mathbf{M}$ is $\mathbf{Q}\mathrm{diag}(\mathrm{sign}(\mathrm{diag}(\mathbf{R})))$, where $\mathrm{sign}(\cdot)$ is the sign function, and $\mathrm{sign}(0) = 1$.

Matrix factorization in backpropagation The matrix-to-vector mapping $\mathcal{F}$ involves eigendecomposition or Cholesky factorization (e.g., $\mathbf{X}_3(t) = \phi^{-1}(\frac{\alpha}{H}\sum_{h=1}^{H}\phi(\mathbf{W}_1^h\mathbf{X}(t)(\mathbf{W}_1^h)^\top))$ for geodesic-shrinkage activated pooling). Computing the partial derivatives required for backpropagation is therefore quite demanding. Appendix E provides the detailed derivations, as well as the functional gradients required for backpropagation from the encoder $\mathcal{E}$ to the mapping $\mathcal{F}$.

5 Experiments

We here compare `MatFAE` with TSRVF (Dai et al., 2020), GeoAtt (Dan et al., 2022), SPDNet (Huang and Gool, 2017), PGA (Fletcher et al., 2004), MLP autoencoder (DC), and k-means applied directly to the pairwise distance matrix (PDM) induced by the LE metric. Appendix G.1 briefly explains each benchmark method and reports the configuration settings used for all algorithms. For methods that produce vector representations, clustering is performed using k-means. The number of clusters is selected by maximizing the silhouette score over $k \in \{2,3,4,5\}$.

Simulation study We consider ten simulation scenarios, and the data-generating mechanism for each is provided in detail in Appendix G.3. For each scenario, we generate 100 independent datasets and apply all benchmark methods to each dataset. Clustering performance is then summarized by the mean and standard deviation of the AMI and ARI across the 100 replications, reported in Tables 1 and 3, respectively. The benchmark is structured as an ablation ladder, with each synthetic dataset designed to isolate a single type of structural signal. Accordingly, failure on a given rung indicates that the method lacks the feature-extraction mechanism needed to capture that particular structure.

Table 1 shows that MatFAE performs best overall across the ten simulation scenarios. It achieves the highest mean AMI in six scenarios (A, C, D, E, F, and I) and ties for the best in H and J. Its main advantage appears in the more difficult settings (C, D, and E), where competing methods show only modest or near-chance performance. The gain in E is particularly pronounced: MatFAE reaches 0.725, whereas the next-best method, PGA, achieves only 0.163.

Real datasets We used six real resting-state fMRI datasets from two public neuroimaging repositories: OpenNeuro and the International Neuroimaging Data-sharing Initiative. Detailed dataset profiles, preprocessing procedures, and additional analyses are provided in Appendix G.4. Appendix G.6 provides a step-by-step demonstration of MatFAE's intrinsic interpretability; it shows how the temporal profiles of the functional weights $\boldsymbol{W}^{(1)}(t)$ identify the portions of the SPD trajectory that contribute most strongly to the latent representations.

We report AMI and ARI in Tables 2 and 4, respectively. `MatFAE` achieves the best performance, demonstrating its consistent advantage over the competing methods. The absolute AMI values however reflect the intrinsic difficulty of unsupervised diagnostic discovery from resting-state fMRI. ADHD-200 and ABIDE-I are heterogeneous multi-site cohorts, CNP contains small patient subgroups across four DSM categories, and resting-state cognition is itself an unconstrained latent state. Even supervised models on these datasets typically improve only moderately over chance, with reported accuracies of approximately 0.60-0.70 for ABIDE-I (Heinsfeld et al., 2018; Saponaro et al., 2022), 0.72-0.81 for COBRE (Zeng et al., 2018), and 0.55-0.66 for ADHD-200 (Qiu et al., 2024).

Classification We further evaluate the encoder blocks of MatFAE in a supervised setting, denoted MatFAE-C, by removing the decoder and appending a classification head to the latent representation. We compare MatFAE-C against three recent classifiers for temporal SPD/graph-structured data on the same six real datasets: SPD-SRU (Chakraborty et al., 2018), STAGIN (Kim et al.,

Table 1: AMI scores for the ten simulation scenarios. The table reports the mean (top row) and standard deviation (bottom row) of the scores over 100 repetitions.

Scenario (*rung*)	TSRVF	GeoAtt	SPDNet	PGA	DC	PDM	MatFAE
A (*temporal ordering*)	0.998	-0.001	-0.000	-0.000	0.989	-0.000	**1.000**
	0.016	0.008	0.010	0.008	0.100	0.009	0.000
B (*static identity*)	0.009	0.021	0.033	**0.998**	0.006	0.949	0.971
	0.025	0.046	0.055	0.012	0.022	0.135	0.013
C (*dwell time*)	0.319	0.447	0.438	0.451	0.444	0.449	**0.593**
	0.064	0.076	0.059	0.063	0.077	0.062	0.040
D (*transition frequency*)	0.350	-0.001	0.000	-0.000	0.336	0.002	**0.368**
	0.069	0.008	0.010	0.010	0.126	0.015	0.044
E (*smoothness*)	0.123	0.115	0.009	0.163	0.062	-0.000	**0.725**
	0.077	0.073	0.024	0.064	0.044	0.000	0.054
F (*Wishart concentration*)	0.125	0.999	0.469	0.753	0.182	0.648	**1.000**
	0.055	0.007	0.067	0.042	0.074	0.100	0.000
G (*real HRF*)	-0.001	0.000	0.003	**1.000**	0.979	0.151	0.929
	0.010	0.011	0.015	0.000	0.119	0.105	0.025
H (*trajectory direction*)	**1.000**	0.149	0.029	0.010	**1.000**	0.002	**1.000**
	0.000	0.075	0.032	0.022	0.000	0.017	0.000
I (*multi-scale dynamics*)	0.141	0.032	0.033	0.031	0.048	0.032	**0.155**
	0.090	0.044	0.043	0.042	0.059	0.042	0.085
J (*Neurosynth*)	0.000	0.235	0.714	**1.000**	0.766	**1.000**	**1.000**
	0.010	0.242	0.134	0.000	0.352	0.000	0.000

Table 2: AMI scores for the six real resting-state fMRI datasets.

Dataset	TSRVF	GeoAtt	SPDNet	PGA	DC	PDM	MatFAE
CNP	0.051	0.015	0.013	0.035	0.015	0.001	**0.109**
COBRE	0.064	0.052	0.052	0.041	0.013	0.041	**0.137**
ADHD-200	0.026	0.003	0.004	0.027	0.053	0.055	**0.078**
ABIDE-I	0.010	0.001	0.001	0.000	0.004	0.001	**0.030**
TCP	0.026	0.002	0.010	0.020	0.004	0.018	**0.042**
CAT-D	0.019	0.022	0.009	0.041	0.003	0.028	**0.062**

2021), and BW-Norm (Wang et al., 2025). Full configuration details for all methods are reported in Appendix G.2. Classification results are reported in Table 5. The supervised variant `MatFAE-C` achieves the highest balanced classification accuracy on all six cohorts. The purpose of `MatFAE-C` is not to claim state-of-the-art supervised diagnostic classification, but to demonstrate that the representation learned by the `MatFAE` encoder contains diagnostically relevant information. In neuroscience, many neurological and psychiatric conditions are heterogeneous and may contain clinically meaningful subtypes that are not known in advance. `MatFAE` can therefore be used to identify latent subgroups and support symptom-driven, transdiagnostic characterizations of psychiatric conditions.

6 Conclusion

We introduced `MatFAE`, a geometry-aware neural architecture for representation learning from functional data valued on the SPD manifold. Beyond combining SPD-aware dimensionality reduction with functional feature extraction, `MatFAE` contributes three theoretically grounded mechanisms: a regularized multi-head BiMap layer that captures heterogeneous, time-varying joint structure inaccessible to a single global projection; a geodesic-shrinkage activation, derived from our analysis of the non-contractive, nonlinear properties of Fréchet-mean pooling, which generalizes to Riemannian manifolds beyond the SPD setting; and an intrinsic interpretability procedure that traces latent representations back to the original SPD-valued trajectories. Together, these components allow `MatFAE` to capture both the geometric structure and the temporal morphology of SPD-valued functional data within a single, lightweight, non-recurrent architecture.

AI USE STATEMENT

A Large language model (LLM) was used for editing the grammar and spelling and as an aid in searching for and identifying relevant literature. All AI-assisted suggested edits were reviewed by the authors, and every literature reference identified with AI assistance was read and verified. All reported experimental results were produced by the authors' own code. We take full responsibility for the content of this work.

ETHICS STATEMENT

This work uses only publicly available, de-identified human neuroimaging datasets released for research use by their respective consortia (OpenNeuro and the International Neuroimaging Data-sharing Initiative), no new data were collected. The proposed method is intended for representation learning research, unsupervised diagnostic discovery from resting-state fMRI remains far from clinical applicability, and the limitations of the real-data analyses are discussed in the Limitations section.

REPRODUCIBILITY STATEMENT

`MatFAE` is available as a fully documented Python package. The anonymous GitHub link is https://anonymous.4open.science/r/MatFAE-main/. All datasets used are publicly available, the preprocessing pipelines, dataset-specific deviations, and model configurations are documented in the Appendix.

REFERENCES

E. Akeweje and M. Zhang. Learning mixtures of Gaussian processes through random projection. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 720–739. PMLR, 21–27 Jul 2024.

E. A. Allen, E. Damaraju, S. M. Plis, E. B. Erhardt, T. Eichele, and V. D. Calhoun. Tracking whole-brain connectivity dynamics in the resting state. *Cerebral Cortex*, 24(3):663–676, 03 2014. doi: 10.1093/cercor/bhs352.

V. Arsigny, P. Fillard, X. Pennec, and N. Ayache. Log-euclidean metrics for fast and simple calculus on diffusion tensors. *Magnetic Resonance in Medicine*, 56(2):411–421, 2006. doi: 10.1002/mrm.20965.

V. Arsigny, P. Fillard, X. Pennec, and N. Ayache. Geometric means in a novel vector space structure on symmetric positive-definite matrices. *SIAM Journal on Matrix Analysis and Applications*, 29 (1):328–347, 2007. doi: 10.1137/050637996.

K. Balasubramanian, H.-G. Müller, and B. K. Sriperumbudur. Functional linear and single-index models: A unified approach via gaussian stein identity. *Bernoulli*, 31(2):973 – 1006, 2025. doi: 10.3150/24-BEJ1755.

P. Bellec, C. Chu, F. Chouinard-Decorte, Y. Benhajali, D. S. Margulies, and R. C. Craddock. The neuro bureau adhd-200 preprocessed repository. *NeuroImage*, 144(Part B):275–286, 2017. doi: 10.1016/j.neuroimage.2016.06.034.

R. Bilder, R. Poldrack, T. Cannon, E. London, N. Freimer, E. Congdon, K. Karlsgodt, and F. Sabb. "ucla consortium for neuropsychiatric phenomics la5c study", 2020.

C. Bonet, B. Malézieux, A. Rakotomamonjy, L. Drumetz, T. Moreau, M. Kowalski, and N. Courty. Sliced-Wasserstein on symmetric positive definite matrices for M/EEG signals. In A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 2777–2805. PMLR, 23–29 Jul 2023.

D. Brooks, O. Schwander, F. Barbaresco, J.-Y. Schneider, and M. Cord. Riemannian batch normalization for SPD neural networks. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

X. Cai, L. Xue, and J. Cao. Variable selection for multiple function-on-function linear regression. *Statistica Sinica*, 32(3):1435 – 1465, 2022. doi: 10.5705/ss.202020.0473.

C. C. Camp, S. Noble, D. Scheinost, A. Stringaris, and D. M. Nielson. Test-retest reliability of functional connectivity in adolescents with depression. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, 9(1):21–29, 2024. doi: 10.1016/j.bpsc.2023.09.002.

R. Chakraborty, C.-H. Yang, X. Zhen, M. Banerjee, D. Archer, D. Vaillancourt, V. Singh, and B. Vemuri. A statistical recurrent model on the manifold of symmetric positive definite matrices. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

R. Chakraborty, J. Bouza, J. H. Manton, and B. C. Vemuri. ManifoldNet: A deep neural network for manifold-valued data with applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(2):799–810, 2022. doi: 10.1109/TPAMI.2020.3003846.

Z. Chen, T. Xu, X.-J. Wu, R. Wang, Z. Huang, and J. Kittler. Riemannian local mechanism for spd neural networks. In *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence and Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence and Thirteenth Symposium on Educational Advances in Artificial Intelligence*. AAAI Press, 2023. doi: 10.1609/aaai.v37i6.25867.

J.-M. Chiou, Y.-T. Chen, and Y.-F. Yang. Multivariate functional principal component analysis: A normalization approach. *Statistica Sinica*, 24(4):1571–1596, 2014.

S. Chopra, C. V. Cocuzza, C. Lawhead, J. A. Ricard, L. Labache, L. M. Patrick, P. Kumar, A. Rubenstein, J. Moses, L. Chen, C. Blankenbaker, B. Gillis, L. T. Germine, I. Harpaz-Rotem, B. T. T. Yeo, J. T. Baker, and A. J. Holmes. The Transdiagnostic Connectome Project: an open dataset for studying brain-behavior relationships in psychiatry. *Scientific Data*, 12(1):923, 2025. doi: 10.1038/s41597-025-04895-z. OpenNeuro `ds005237`.

M. Dai, Z. Zhang, and A. Srivastava. Analyzing dynamical brain functional connectivity as trajectories on space of covariance matrices. *IEEE Transactions on Medical Imaging*, 39(3):611–620, 2020. doi: 10.1109/TMI.2019.2931708.

T. Dan, Z. Huang, H. Cai, P. J. Laurienti, and G. Wu. Learning brain dynamics of evolving manifold functional mri data using geometric-attention neural network. *IEEE Transactions on Medical Imaging*, 41(10):2752–2763, 2022. doi: 10.1109/TMI.2022.3169640.

H. Dette and J. Tang. Statistical inference for function-on-function linear regression. *Bernoulli*, 30 (1):304 – 331, 2024. doi: 10.3150/23-BEJ1598.

A. Di Martino, C.-G. Yan, Q. Li, E. Denio, F. X. Castellanos, K. Alaerts, J. S. Anderson, M. Assaf, S. Y. Bookheimer, M. Dapretto, B. Deen, S. Delmonte, I. Dinstein, B. Ertl-Wagner, D. A. Fair, L. Gallagher, D. P. Kennedy, C. L. Keown, C. Keysers, J. E. Lainhart, C. Lord, B. Luna, V. Menon, N. J. Minshew, C. S. Monk, S. Mueller, R.-A. Müller, M. B. Nebel, J. T. Nigg, K. O'Hearn, K. A. Pelphrey, S. J. Peltier, J. D. Rudie, S. Sunaert, M. Thioux, J. M. Tyszka, L. Q. Uddin, J. S. Verhoeven, N. Wenderoth, J. L. Wiggins, S. H. Mostofsky, and M. P. Milham. The autism brain imaging data exchange: towards a large-scale evaluation of the intrinsic brain architecture in autism. *Molecular Psychiatry*, 19(6):659–667, 2014. doi: 10.1038/mp.2013.78.

O. Esteban, C. J. Markiewicz, R. W. Blair, C. A. Moodie, A. I. Isik, A. Erramuzpe, J. D. Kent, M. Goncalves, E. DuPre, M. Snyder, H. Oya, S. S. Ghosh, J. Wright, J. Durnez, R. A. Poldrack, and K. J. Gorgolewski. fMRIPrep: a robust preprocessing pipeline for functional MRI. *Nature Methods*, 16(1):111–116, 2019. doi: 10.1038/s41592-018-0235-4.

K. Fan and G. Pall. Imbedding conditions for Hermitian and normal matrices. *Canadian Journal of Mathematics*, 9:298 – 304, 1957. doi: 10.4153/CJM-1957-036-1.

S. Fine, Y. Singer, and N. Tishby. The hierarchical hidden markov model: Analysis and applications. *Machine Learning*, 32(1):41 – 62, 1998. doi: 10.1023/A:1007469218079.

P. Fletcher, C. Lu, S. Pizer, and S. Joshi. Principal geodesic analysis for the study of nonlinear statistics of shape. *IEEE Transactions on Medical Imaging*, 23(8):995–1005, 2004. doi: 10.1109/TMI.2004.831793.

M. F. Glasser, T. S. Coalson, E. C. Robinson, C. D. Hacker, J. Harwell, E. Yacoub, K. Ugurbil, J. Andersson, C. F. Beckmann, M. Jenkinson, S. M. Smith, and D. C. Van Essen. A multimodal parcellation of human cerebral cortex. *Nature*, 536(7615):171–178, 2016. doi: 10.1038/nature18933.

G. H. Glover. Deconvolution of impulse response in event-related bold fmri1. *NeuroImage*, 9(4): 416–429, 1999. doi: https://doi.org/10.1006/nimg.1998.0419.

K. J. Gorgolewski, J. Durnez, and R. A. Poldrack. Preprocessed Consortium for Neuropsychiatric Phenomics dataset. *F1000Research*, 6:1262, 2017. doi: 10.12688/f1000research.11964.2.

K. Guo, P. Ishwar, and J. Konrad. Action recognition from video using feature covariance matrices. *IEEE Transactions on Image Processing*, 22(6):2479–2494, 2013. doi: 10.1109/TIP.2013.2252622.

P. Hall and J. L. Horowitz. Methodology and convergence rates for functional linear regression. *Annals of Statistics*, 35(1):70 – 91, 2007. doi: 10.1214/009053606000000957.

F. Heinrichs, M. Heim, and C. Weber. Functional neural networks: Shift invariant models for functional data with applications to EEG classification. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 12866–12881. PMLR, 23–29 Jul 2023.

A. S. Heinsfeld, A. R. Franco, R. C. Craddock, A. Buchweitz, and F. Meneguzzi. Identification of autism spectrum disorder using deep learning and the ABIDE dataset. *NeuroImage: Clinical*, 17: 16–23, 2018. doi: 10.1016/j.nicl.2017.08.017.

Z. Huang and L. V. Gool. A Riemannian network for SPD matrix learning. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, page 2036–2042. AAAI Press, 2017.

Z. Huang, R. Wang, S. Shan, X. Li, and X. Chen. Log-euclidean metric learning on symmetric positive definite manifold with application to image set classification. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*, page 720–729. JMLR.org, 2015.

C. Jang, Y. Lee, Y.-K. Noh, and F. C. Park. Geometrically regularized autoencoders for non-euclidean data. In *The Eleventh International Conference on Learning Representations*, 2023.

B.-H. Kim, J. C. Ye, and J.-J. Kim. Learning dynamic graph representation of brain connectome with spatio-temporal attention. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 4314–4327. Curran Associates, Inc., 2021.

Z. Lin. Riemannian geometry of symmetric positive definite matrices via Cholesky decomposition. *SIAM Journal on Matrix Analysis and Applications*, 40(4):1353–1370, 2019. doi: 10.1137/18M1221084.

H. Liu, J. Li, Y. Wu, and R. Ji. Learning neural bag-of-matrix-summarization with riemannian network. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence*. AAAI Press, 2019. doi: 10.1609/aaai.v33i01.33018746.

F. López, B. Pozzetti, S. Trettel, M. Strube, and A. Wienhard. Vector-valued distance and gyrocalculus on the space of symmetric positive definite matrices. In *Proceedings of the 35th International Conference on Neural Information Processing Systems*, NIPS '21. Curran Associates Inc., 2021.

D. J. Lurie, D. Kessler, D. S. Bassett, R. F. Betzel, M. Breakspear, S. Keilholz, A. Kucyi, R. Liégeois, M. A. Lindquist, A. R. McIntosh, R. A. Poldrack, J. M. Shine, W. H. Thompson, N. Z. Bielczyk, L. Douw, D. Kraft, R. L. Miller, M. Muthuraman, L. Pasquini, A. Razi, D. Vidaurre, H. Xie, and V. D. Calhoun. Questions and controversies in the study of time-varying functional connectivity in resting fMRI. *Network Neuroscience*, 4(1):30 – 69, 2020. doi: 10.1162/netn_a_00116.

X. S. Nguyen, Y. Yang, and A. Histace. Matrix manifold neural networks++. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun, editors, *International Conference on Learning Representations*, volume 2024, pages 40448–40478, 2024.

V. Patilea, C. Sánchez-Sellero, and M. Saumard. Testing the predictor effect on a functional response. *Journal of the American Statistical Association*, 111(516):1684–1695, 2016. doi: 10.1080/01621459.2015.1110031.

X. Pennec, P. Fillard, and N. Ayache. A Riemannian framework for tensor computing. *International Journal of Computer Vision*, 66(1):41 – 66, 2006. doi: 10.1007/s11263-005-3222-z.

B. Qiu, Q. Wang, X. Li, W. Li, W. Shao, and M. Wang. Adaptive spatial-temporal neural network for ADHD identification using functional fMRI. *Frontiers in Neuroscience*, 18:1394234, 2024. doi: 10.3389/fnins.2024.1394234.

F. Rossi, B. Conan-Guez, and F. Fleuret. Functional data analysis with multi layer perceptrons. In *2002 International Joint Conference on Neural Networks (IJCNN)*, volume 3, pages 2843–2848 vol.3, 2002. doi: 10.1109/IJCNN.2002.1007599.

N. Sadeghi, P. Q. Fors, L. Eisner, J. Taigman, K. Qi, L. S. Gorham, C. C. Camp, G. O'Callaghan, D. Rodriguez, J. McGuire, E. M. Garth, C. Engel, M. Davis, K. E. Towbin, A. Stringaris, and D. M. Nielson. Mood and behaviors of adolescents with depression in a longitudinal study before and during the COVID-19 pandemic. *Journal of the American Academy of Child & Adolescent Psychiatry*, 61(11):1341–1350, 2022. doi: 10.1016/j.jaac.2022.04.004. OpenNeuro `ds004627`.

G. Salimi-Khorshidi, G. Douaud, C. F. Beckmann, M. F. Glasser, L. Griffanti, and S. M. Smith. Automatic denoising of functional MRI data: combining independent component analysis and hierarchical fusion of classifiers. *NeuroImage*, 90:449–468, 2014. doi: 10.1016/j.neuroimage.2013.11.046.

T. Salo, T. Yarkoni, T. E. Nichols, J.-B. Poline, M. Bilgel, K. L. Bottenhorn, S. B. Eickhoff, D. Jarecka, J. D. Kent, A. Kimbler, D. M. Nielson, K. M. Oudyk, J. A. Peraza, A. Pérez, P. C. Reeders, J. A. Yanes, and A. R. Laird. NiMARE: Neuroimaging Meta-Analysis Research Environment. *Aperture Neuro*, 3:1–32, 2023. doi: 10.52294/001c.87681.

X. Sanchez-Vila, A. Guadagnini, and J. Carrera. Representative hydraulic conductivities in saturated groundwater flow. *Reviews of Geophysics*, 44(3), 2006. doi: 10.1029/2005RG000169.

S. Saponaro, A. Giuliano, R. Bellotti, A. Lombardi, S. Tangaro, P. Oliva, S. Calderoni, and A. Retico. Multi-site harmonization of MRI data uncovers machine-learning discrimination capability in barely separable populations: An example from the ABIDE dataset. *NeuroImage: Clinical*, 35:103082, 2022. doi: 10.1016/j.nicl.2022.103082.

S. Singh, S. Coyle, and M. Zhang. Shape-informed clustering of multi-dimensional functional data via deep functional autoencoders. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2025.

The ADHD-200 Consortium. The adhd-200 consortium: A model to advance the translational potential of neuroimaging in clinical neuroscience. *Frontiers in Systems Neuroscience*, 6:62, 2012. doi: 10.3389/fnsys.2012.00062.

B. Thind, K. Multani, and J. Cao. Deep learning with functional inputs. *Journal of Computational and Graphical Statistics*, 0(0):1–10, 2022. doi: 10.1080/10618600.2022.2097914.

O. Tuzel, F. Porikli, and P. Meer. Pedestrian detection via classification on riemannian manifolds. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(10):1713–1727, 2008. doi: 10.1109/TPAMI.2008.75.

D. Vidaurre, S. M. Smith, and M. W. Woolrich. Brain network dynamics are hierarchically organized in time. *Proceedings of the National Academy of Sciences*, 114(48):12827–12832, 2017. doi: 10.1073/pnas.1705120114.

J.-L. Wang, J.-M. Chiou, and H.-G. Müller. Functional data analysis. *Annual Review of Statistics and Its Application*, 3(Volume 3, 2016):257–295, 2016. doi: 10.1146/annurev-statistics-041715-033624.

R. Wang, S. Jin, Z. Chen, X. Luo, and X.-J. Wu. Learning to normalize on the spd manifold under bures-wasserstein geometry. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8289–8298, June 2025.

T. Yarkoni, R. A. Poldrack, T. E. Nichols, D. C. Van Essen, and T. D. Wager. Large-scale automated synthesis of human functional neuroimaging data. *Nature Methods*, 8(8):665–670, 2011. doi: 10.1038/nmeth.1635.

H. Yeon, X. Dai, and D. J. Nordman. Bootstrap inference in functional linear regression models with scalar response. *Bernoulli*, 29(4):2599 – 2626, 2023. doi: 10.3150/22-BEJ1554.

Y. Yuan, H. Zhu, M. Styner, J. H. Gilmore, and J. Marron. Varying coefficient model for modeling diffusion tensors along white matter tracts. *Annals of Applied Statistics*, 7(1):102 – 125, 2013. doi: 10.1214/12-AOAS574.

L.-L. Zeng, H. Wang, P. Hu, B. Yang, W. Pu, H. Shen, X. Chen, Z. Liu, H. Yin, Q. Tan, K. Wang, and D. Hu. Multi-site diagnostic classification of schizophrenia using discriminant deep learning with functional connectivity MRI. *EBioMedicine*, 30:74–85, 2018. doi: 10.1016/j.ebiom.2018.03.017.

M. Zhang and A. Parnell. Review of clustering methods for functional data. *ACM Transactions on Knowledge Discovery from Data*, 17(7):34, 2023. doi: 10.1145/3581789.

T. Zhang, W. Zheng, Z. Cui, and C. Li. Deep manifold-to-manifold transforming network. In *2018 25th IEEE International Conference on Image Processing (ICIP)*, pages 4098–4102, 2018. doi: 10.1109/ICIP.2018.8451626.

A Related work

Functional data analysis Functional data analysis studies data whose basic observational units are functions (e.g., curves over time or space) and, more broadly, other complex objects such as images and shapes. In functional data analysis, each subject in a sample is represented by one or multiple functions observed over a continuum (such as time, wavelength, or spatial location). Let $\{\boldsymbol{y}_i : i = 1, \ldots, n\}$ be a set of n independent p-dimensional sample functions: $\boldsymbol{y}_i(t) = (y_i^1(t), \ldots, y_i^p(t))^\top \in \mathbb{R}^p$ for any $t \in [t_0, t_1]$; the superscript $\top$ is the transpose operator. Each sample function $\boldsymbol{y}_i$ is a random realization of a p-dimensional random function $\vec{Y} = (Y^1, \ldots, Y^p)^\top$. For $d = 1, \ldots, p$, the n one-dimensional sample functions $\{y_1^d, \ldots, y_n^d\}$ are independent realizations of the component random function $Y^d = Y^d(t) = Y^d(t, \omega)$, defined on a probability space $(\Omega, \mathcal{F}, \Pr)$ and taking values in $\mathcal{H}([t_0, t_1], \mathbb{R})$. Here, $\mathcal{H}([t_0, t_1], \mathbb{R})$ is the separable Hilbert space of all square-integrable measurable functions that are defined on $[t_0, t_1]$ and taking values in $\mathbb{R}$. That is, Y^d is a measurable map from Ω to $\mathcal{H}([t_0, t_1], \mathbb{R})$.

Functional regression is commonly classified by whether the response and the covariates are functional or finite-dimensional (vector/scalar). This yields three main settings: (i) functional responses with functional covariates Dette and Tang (2024); Cai et al. (2022), (ii) vector (including scalar) responses with functional covariates Hall and Horowitz (2007); Yeon et al. (2023); Balasubramanian et al. (2025), and (iii) functional responses with vector covariates Patilea et al. (2016). Among these, the most extensively studied case is (ii), particularly scalar-on-function regression, where the response is a scalar and the predictor is a function. For an overview of statistical tools in functional data analysis, see Wang et al. (2016). Beyond regression, clustering is a core task in functional data analysis. Zhang and Parnell (2023) provided a comprehensive review of clustering methods for functional data and observed that approaches for clustering multi-dimensional functional data generally follow one of two paradigms: (1) first applying multivariate functional principal component analysis Chiou et al. (2014) and then clustering the resulting score vectors using standard multivariate clustering methods; or (2) defining an appropriate (dis)similarity measure for multi-dimensional functional data and clustering directly using distance- or similarity-based algorithms.

More recently, functional neural networks have been proposed to capture nonlinear relationships that may be difficult to represent with classical functional data analysis models Rossi et al. (2002); Thind et al. (2022); Heinrichs et al. (2023). Singh et al. (2025) provided the first proof that functional neural network has the universal approximation property. A key distinction of these architectures is the inclusion of functional layers, in which the learnable weights are themselves functions rather than scalar parameters, enabling the architectures operate directly on functional inputs. Although one can treat the sequence of discrete evaluations of each $y_i^d(t)$ over a grid as a time series, enabling the use of recurrent neural networks or Transformer-based models, such approaches do not naturally encode shape (morphological) characteristics or derivative-based (first-order) information of the underlying functions. By contrast, functional neural networks offer a principled framework for learning representations that respect the functional nature of the data rather than relying solely on discrete measurements.

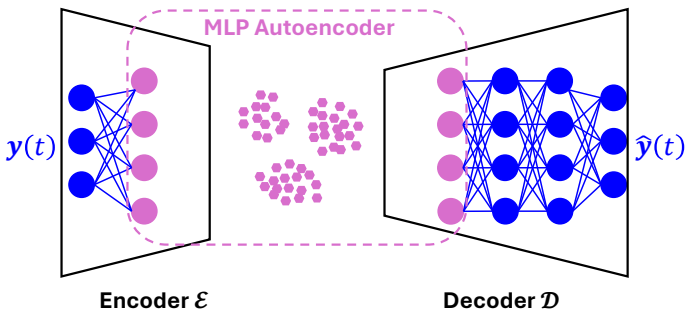

Figure 2: The FAEclust architecture generalizes the multilayer perceptron (MLP) autoencoder by appending functional layers at the encoder input and the decoder output. In the above figure, the plum block denotes the MLP autoencoder, and the blue components denote functional nodes/edges.

Riemannian neural networks Deep neural networks for SPD matrices must respect the fact that the data live on a Riemannian manifold. Over the past decade, two main modeling paradigms have emerged: tangent-space learning and intrinsic manifold deep networks, with more recent work extending these ideas to attention, normalization, and alternative geometries. In the tangent-space learning paradigm, we map an SPD matrix to a vector space using a Riemannian logarithm map, then apply standard neural architectures on the mapped data Arsigny et al. (2006); Tuzel et al. (2008); Huang et al. (2015). This approach is computationally attractive because it avoids repeated eigen/SVD constraints inside the network while keeping a principled geometric link to the manifold. In the intrinsic deep learning paradigm, every operation preserves positive definiteness. The canonical example is SPDNet Huang and Gool (2017), which introduced a stack of SPD-preserving layers: bilinear mapping (BiMap) for dimension reduction on SPD, eigenvalue rectification (ReEig) as a nonlinearity, and log-eigenvalue (LogEig) to move to a Euclidean space only near the output for standard classifiers. This architecture family has been extended in multiple directions: regularization on manifolds Jang et al. (2023), Riemannian batch normalization on the SPD manifold Brooks et al. (2019); Wang et al. (2025), a multi-channel mechanism to capture multi-scale features Chen et al. (2023), and replacing LogEig with a learnable SPD encoder Liu et al. (2019). Another emerging direction is to broaden the layer toolbox (e.g., convolutional layers) to make SPD networks more closely aligned with standard deep learning building blocks, while still ensuring valid manifold backpropagation Nguyen et al. (2024).

To the best of our knowledge, there is currently no deep neural network architecture specifically designed to model SPD-valued trajectories $\{\mathbf{X}_i \in \mathcal{H}_g([t_0, t_1]; \mathcal{S}_+^m)\}_{i=1}^n$, where each observation $\mathbf{X}_i$ is a function of time, and its value $\mathbf{X}_i(t)$ is an SPD matrix. Such SPD-valued functional data arise naturally when second-order statistics evolve continuously over time, and they are increasingly encountered in modern data-rich scientific domains, most notably neuroscience. A representative example is magnetoencephalography (MEG) or electroencephalography (EEG) analysis. Let $\boldsymbol{x}_i(t) \in \mathbb{R}^m$ denote the multichannel sensor signal of subject i at time t. Instead of analyzing raw signals, it is common to compute time-resolved covariance or cross-spectral matrices $\mathbf{X}_i(t)$ over short, possibly overlapping, temporal windows. As the window slides continuously, this procedure yields a smooth trajectory of SPD matrices over time, capturing the dynamic evolution of brain networks during cognitive tasks, motor imagery, or resting-state activity Bonet et al. (2023). In this work, we develop `MatFAE` to bridge this methodological gap by unifying recent advances in Riemannian deep learning and functional data analysis for end-to-end modeling of SPD-valued trajectories.

B Properties of the two Riemannian manifolds

B.1 The Log-Cholesky metric

Let $\mathcal{L}^m$ denote the set of $m \times m$ lower triangular matrices, and $\mathcal{L}_+^m \subset \mathcal{L}^m$ the subset of lower triangular matrices with positive diagonal elements. Both $\mathcal{L}_+^m$ and $\mathcal{S}_+^m$ are smooth manifolds and therefore have differential geometry properties. The tangent space $\mathcal{T}_{\mathbf{L}}\mathcal{L}_+^m$ of $\mathcal{L}_+^m$ at any point $\mathbf{L} \in \mathcal{L}_+^m$ is $\mathcal{L}^m$. For any $\mathbf{X} \in \mathcal{S}_+^m$, there is a unique lower triangular matrix with positive diagonal elements $\mathbf{L} = \mathfrak{L}(\mathbf{X}) \in \mathcal{L}_+^m$ that satisfies $\mathbf{X} = \mathbf{L}\mathbf{L}^\top$. Furthermore, the map $\mathfrak{L} : \mathcal{S}_+^m \to \mathcal{L}_+^m$ is a diffeomorphism (Lin, 2019, Proposition 2). The differential $D_{\mathbf{X}}\mathfrak{L} : \mathcal{T}_{\mathbf{X}}\mathcal{S}_+^m \mapsto \mathcal{T}_{\mathfrak{L}(\mathbf{X})}\mathcal{L}_+^m$ of $\mathfrak{L}$ at $\mathbf{X}$ is given by

$$(D_{\mathbf{X}}\mathfrak{L})(\mathbf{U}) = \mathfrak{L}(\mathbf{X})(\mathfrak{L}(\mathbf{X})^{-1}\mathbf{U}\mathfrak{L}(\mathbf{X})^{-\top})_{\frac{1}{2}}, \ \ \mathbf{U} \in \mathcal{T}_{\mathbf{X}}\mathcal{S}_+^m,$$

where $(\mathbf{V})_{\frac{1}{2}} = \lfloor \mathbf{V} \rfloor + \frac{1}{2}\mathrm{diag}(\mathbf{V})$ is the lower triangular part of $\mathbf{V}$ with the diagonal elements halved, and $\mathrm{diag}(\mathbf{V})$ is the matrix of the main diagonal elements of $\mathbf{V}$. Let $\mathfrak{L}^{-1} : \mathcal{L}_+^m \mapsto \mathcal{S}_+^m$ denote the inverse of the map $\mathfrak{L}$: $\mathfrak{L}^{-1}(\mathbf{L}) = \mathbf{L}\mathbf{L}^\top$. The differential $D_{\mathbf{L}}\mathfrak{L}^{-1} : \mathcal{T}_{\mathbf{L}}\mathcal{L}_+^m \mapsto \mathcal{T}_{\mathbf{L}\mathbf{L}^\top}\mathcal{S}_+^m$ of $\mathfrak{L}^{-1}$ at $\mathbf{L}$ is given by

$$(D_{\mathbf{L}}\mathfrak{L}^{-1})(\mathbf{U}) = \mathbf{L}\mathbf{U}^\top + \mathbf{U}\mathbf{L}^\top, \ \ \mathbf{U} \in \mathcal{T}_{\mathbf{L}}\mathcal{L}_+^m.$$

Because $\mathfrak{L}$ is a diffeomorphism, the Riemannian metric for $\mathcal{S}_+^m$ can be induced (via pullback) from the Riemannian metric for $\mathcal{L}_+^m$.

For the manifold $\mathcal{L}_+^m$, the Riemannian metric at a point $\mathbf{L} \in \mathcal{L}_+^m$ is

$$\langle \mathbf{U}, \mathbf{V} \rangle_{\mathbf{L}}^{\mathcal{L}_+^m} = \langle \lfloor \mathbf{U} \rfloor, \lfloor \mathbf{V} \rfloor \rangle_F + \langle \mathrm{diag}(\mathbf{L})^{-1}\mathrm{diag}(\mathbf{U}), \mathrm{diag}(\mathbf{L})^{-1}\mathrm{diag}(\mathbf{V}) \rangle_F, \ \ \mathbf{U}, \mathbf{V} \in \mathcal{T}_{\mathbf{L}}\mathcal{L}_+^m,$$

where $\langle \cdot, \cdot \rangle_F$ is the Frobenius inner product, with induced norm $\| \cdot \|_F$. The Riemannian exponential map at $\mathbf{L}$ is

$$\mathrm{Exp}_{\mathbf{L}}(\mathbf{U}) = \lfloor \mathbf{L} \rfloor + \lfloor \mathbf{U} \rfloor + \mathrm{diag}(\mathbf{L}) \exp(\mathrm{diag}(\mathbf{U}) \mathrm{diag}(\mathbf{L})^{-1}), \;\; \mathbf{U} \in \mathcal{T}_{\mathbf{L}} \mathcal{L}_+^m,$$

and the Riemannian logarithmic map at $\mathbf{L}$ is

$$\mathrm{Log}_{\mathbf{L}}(\mathbf{K}) = \lfloor \mathbf{K} \rfloor - \lfloor \mathbf{L} \rfloor + \mathrm{diag}(\mathbf{L}) \log(\mathrm{diag}(\mathbf{L})^{-1} \mathrm{diag}(\mathbf{K})), \;\; \mathbf{K} \in \mathcal{L}_+^m.$$

The geodesic distance between two $\mathbf{L}$ and $\mathbf{K} \in \mathcal{L}_+^m$ is

$$d_{\mathcal{L}_+^m}(\mathbf{L}, \mathbf{K}) = \sqrt{\| \lfloor \mathbf{L} \rfloor - \lfloor \mathbf{K} \rfloor \|_F^2 + \| \log(\mathrm{diag}(\mathbf{L})) - \log(\mathrm{diag}(\mathbf{K})) \|_F^2}.$$

Riemannian geometry Given the Riemannian metric $\langle \cdot, \cdot \rangle_{\cdot}^{\mathcal{L}_+^m}$ for $\mathcal{L}_+^m$, the manifold map $\mathfrak{L} : \mathcal{S}_+^m \mapsto \mathcal{L}_+^m$ induces the Riemannian metric for $\mathcal{S}_+^m$, called the Log-Cholesky (LC) metric:

$$\langle \mathbf{U}, \mathbf{V} \rangle_{\mathbf{X}}^{\mathrm{LC}} = \langle (D_{\mathbf{X}} \mathfrak{L})(\mathbf{U}), (D_{\mathbf{X}} \mathfrak{L})(\mathbf{V}) \rangle_{\mathfrak{L}(\mathbf{X})}^{\mathcal{L}_+^m}, \;\; \mathbf{U}, \mathbf{V} \in \mathcal{T}_{\mathbf{X}} \mathcal{S}_+^m.$$

The Riemannian exponential map at $\mathbf{X} \in \mathcal{S}_+^m$ is

$$\mathrm{Exp}_{\mathbf{X}}^{\mathrm{LC}}(\mathbf{V}) = [\mathrm{Exp}_{\mathfrak{L}(\mathbf{X})}((D_{\mathbf{X}} \mathfrak{L})(\mathbf{V}))][\mathrm{Exp}_{\mathfrak{L}(\mathbf{X})}((D_{\mathbf{X}} \mathfrak{L})(\mathbf{V}))]^\top, \;\; \mathbf{V} \in \mathcal{T}_{\mathbf{X}} \mathcal{S}_+^m.$$

The Riemannian logarithmic map at $\mathbf{X}$ is

$$\mathrm{Log}_{\mathbf{X}}^{\mathrm{LC}}(\mathbf{Z}) = (D_{\mathfrak{L}(\mathbf{X})} \mathfrak{L}^{-1})(\mathrm{Log}_{\mathfrak{L}(\mathbf{X})}(\mathfrak{L}(\mathbf{Z}))), \;\; \mathbf{Z} \in \mathcal{S}_+^m. \tag{5}$$

When the reference point is the identity matrix $\mathbf{I}$, the Riemannian logarithmic map of $\mathbf{Z}$ is

$$\mathrm{Log}_{\mathbf{I}}^{\mathrm{LC}}(\mathbf{Z}) = \lfloor \mathfrak{L}(\mathbf{Z}) \rfloor + \lfloor \mathfrak{L}(\mathbf{Z}) \rfloor^\top + 2 \log(\mathrm{diag}(\mathfrak{L}(\mathbf{Z}))). \tag{6}$$

The map $\mathfrak{L}$ is an isometry between $(\mathcal{S}_+^m, \langle \cdot, \cdot \rangle_{\cdot}^{\mathrm{LC}})$ and $(\mathcal{L}_+^m, \langle \cdot, \cdot \rangle_{\cdot}^{\mathcal{L}_+^m})$, and therefore the geodesic distance between $\mathbf{X}, \mathbf{Z} \in \mathcal{S}_+^m$ is simply $d_{\mathrm{LC}}(\mathbf{X}, \mathbf{Z}) = d_{\mathcal{L}_+^m}(\mathfrak{L}(\mathbf{X}), \mathfrak{L}(\mathbf{Z}))$.

Fréchet mean Given a set of SPD matrices $\{\mathbf{X}_i \in \mathcal{S}_+^m : i = 1, \ldots, n\}$, under mild conditions, the Fréchet mean $\mathbb{E}_{\mathrm{LC}}(\mathbf{X}_{1:n})$ exists and is unique:

$$\mathbb{E}_{\mathrm{LC}}(\mathbf{X}_{1:n}) = \operatorname*{arg\,min}_{\mathbf{X} \in \mathcal{S}_+^m} \sum_{i=1}^n d_{\mathrm{LC}}(\mathbf{X}, \mathbf{X}_i)^2 = \mathbf{L}_n \mathbf{L}_n^\top,$$

where

$$\mathbf{L}_n = \frac{1}{n} \sum_{i=1}^n \lfloor \mathfrak{L}(\mathbf{X}_i) \rfloor + \exp(\frac{1}{n} \sum_{i=1}^n \log(\mathrm{diag}(\mathfrak{L}(\mathbf{X}_i)))). \tag{7}$$

The sectional curvature of $(\mathcal{S}_+^m, \langle \cdot, \cdot \rangle_{\cdot}^{\mathrm{LC}})$ is constantly zero (Lin, 2019, Proposition 8), and therefore $(\mathcal{S}_+^m, \langle \cdot, \cdot \rangle_{\cdot}^{\mathrm{LC}})$ is a Cartan-Hadamard manifold.

B.2 The Log-Euclidean metric

Every real symmetric matrix $\mathbf{U} \in \mathrm{Sym}(m)$ admits an orthogonal eigen-decomposition: $\mathbf{U} = \mathbf{Q} \mathrm{diag}(\lambda_1, \ldots, \lambda_m) \mathbf{Q}^\top$, and therefore taking the matrix exponential can be done by simply applying its scalar version to eigenvalues:

$$\exp(\mathbf{U}) = \mathbf{Q} \mathrm{diag}(\exp(\lambda_1), \ldots, \exp(\lambda_m)) \mathbf{Q}^\top.$$

Furthermore, for an SPD matrix $\mathbf{X} \in \mathcal{S}_+^m \subset \mathrm{Sym}(m)$, the (principal) matrix logarithm exists, is unique, and also admits the eigen-mapping form:

$$\log(\mathbf{X}) = \mathbf{Q} \mathrm{diag}(\log(\gamma_1), \ldots, \log(\gamma_m)) \mathbf{Q}^\top,$$

where $\mathbf{X} = \mathbf{Q} \mathrm{diag}(\gamma_1, \ldots, \gamma_m) \mathbf{Q}^\top$ is the eigen-decomposition. Both $\exp : \mathrm{Sym}(m) \mapsto \mathcal{S}_+^m$ and its inverse $\log : \mathcal{S}_+^m \mapsto \mathrm{Sym}(m)$ are diffeomorphisms. Let $(D_{\mathbf{X}} \log)(\mathbf{U})$ be the differential (Fréchet derivative) of the matrix logarithm at the reference point $\mathbf{X}$, evaluated in the direction $\mathbf{U}$, and $(D_{\mathbf{U}} \exp)(\mathbf{V})$ the differential of the matrix exponential.

Let $\mathbf{X} = \mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^\top$ with $\mathbf{\Lambda} = \mathrm{diag}(\gamma_1, \ldots, \gamma_m)$. The differential $D_{\mathbf{X}} \log : \mathcal{T}_{\mathbf{X}}\mathcal{S}_+^m \mapsto \mathcal{T}_{\log(\mathbf{X})}\mathrm{Sym}(m)$ of the matrix logarithm at $\mathbf{X}$ is given by

$$(D_{\mathbf{X}} \log)(\mathbf{U}) = \mathbf{Q}(\hat{\mathbf{\Lambda}} \odot \tilde{\mathbf{U}})\mathbf{Q}^\top, \;\; \mathbf{U} \in \mathcal{T}_{\mathbf{X}}\mathcal{S}_+^m, \tag{8}$$

where $\odot$ is the Hadamard product, $\tilde{\mathbf{U}} = \mathbf{Q}^\top\mathbf{U}\mathbf{Q}$, and the Loewner matrix $\hat{\mathbf{\Lambda}}$ has diagonal entries $\hat{\Lambda}_{i,i} = \gamma_i^{-1}$ and off-diagonal entries $\hat{\Lambda}_{i,j} = \frac{\log(\gamma_i) - \log(\gamma_j)}{\gamma_i - \gamma_j}$.

Let $\mathbf{U} = \mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^\top$ with $\mathbf{\Lambda} = \mathrm{diag}(\lambda_1, \ldots, \lambda_m)$. The differential $D_{\mathbf{U}} \exp : \mathcal{T}_{\mathbf{U}}\mathrm{Sym}(m) \mapsto \mathcal{T}_{\exp(\mathbf{U})}\mathcal{S}_+^m$ of matrix exponential map at $\mathbf{U}$ is given by

$$(D_{\mathbf{U}} \exp)(\mathbf{K}) = \mathbf{Q}(\hat{\mathbf{\Lambda}} \odot \tilde{\mathbf{K}})\mathbf{Q}^\top, \;\; \mathbf{V} \in \mathcal{T}_{\mathbf{U}}\mathrm{Sym}(m), \tag{9}$$

where $\tilde{\mathbf{K}} = \mathbf{Q}^\top\mathbf{K}\mathbf{Q}$, and the Loewner matrix $\hat{\mathbf{\Lambda}}$ has diagonal entries $\hat{\Lambda}_{i,i} = \exp(\lambda_i)$ and off-diagonal entries $\hat{\Lambda}_{i,j} = \frac{\exp(\lambda_i) - \exp(\lambda_j)}{\lambda_i - \lambda_j}$.

Riemannian geometry The LE Riemannian metric for $\mathcal{S}_+^m$ at $\mathbf{X} \in \mathcal{S}_+^m$ is

$$\langle \mathbf{U}, \mathbf{V} \rangle_{\mathbf{X}}^{\mathrm{LE}} = \langle (D_{\mathbf{X}} \log)(\mathbf{U}), (D_{\mathbf{X}} \log)(\mathbf{V}) \rangle_F, \;\; \mathbf{U}, \mathbf{V} \in \mathcal{T}_{\mathbf{X}}\mathcal{S}_+^m.$$

The Riemannian exponential map at $\mathbf{X} \in \mathcal{S}_+^m$ is

$$\mathrm{Exp}_{\mathbf{X}}^{\mathrm{LE}}(\mathbf{V}) = \exp(\log(\mathbf{X}) + (D_{\mathbf{X}} \log)(\mathbf{V})), \;\; \mathbf{V} \in \mathcal{T}_{\mathbf{X}}\mathcal{S}_+^m.$$

The Riemannian logarithmic map at $\mathbf{X}$ is

$$\mathrm{Log}_{\mathbf{X}}^{\mathrm{LE}}(\mathbf{Z}) = (D_{\log(\mathbf{X})} \exp)(\log(\mathbf{Z}) - \log(\mathbf{X})), \;\; \mathbf{Z} \in \mathcal{S}_+^m. \tag{10}$$

When the reference point is the identity matrix $\mathbf{I}$, the matrix logarithm $\log(\mathbf{I})$ is the zero matrix, and hence the differential $D_{\log(\mathbf{X})} \exp$ is the identity linear map. Therefore, we have $\mathrm{Log}_{\mathbf{I}}^{\mathrm{LE}}(\mathbf{Z}) = \log(\mathbf{Z})$. The matrix exponential map $\exp$ is an isometry between $(\mathcal{S}_+^m, \langle \cdot, \cdot \rangle_\cdot^{\mathrm{LE}})$ and $(\mathrm{Sym}(m), \langle \cdot, \cdot \rangle_F)$, and therefore the geodesic distance between $\mathbf{X}, \mathbf{Z} \in \mathcal{S}_+^m$ is simply $d_{\mathrm{LE}}(\mathbf{X}, \mathbf{Z}) = \| \log(\mathbf{X}) - \log(\mathbf{Z}) \|_F$.

Fréchet mean Given a set of SPD matrices $\{\mathbf{X}_i \in \mathcal{S}_+^m : i = 1, \ldots, n\}$, the Fréchet mean $\mathbb{E}_{\mathrm{LE}}(\mathbf{X}_{1:n})$ exists, is unique, and is simply the arithmetic mean in the domain of matrix logarithms:

$$\mathbb{E}_{\mathrm{LE}}(\mathbf{X}_{1:n}) = \operatorname*{arg\,min}_{\mathbf{X} \in \mathcal{S}_+^m} \sum_{i=1}^n d_{\mathrm{LE}}(\mathbf{X}, \mathbf{X}_i)^2 = \exp(\frac{1}{n} \sum_{i=1}^n \log(\mathbf{X}_i)). \tag{11}$$

Under the LE metric, $(\mathcal{S}_+^m, \langle \cdot, \cdot \rangle_\cdot^{\mathrm{LE}})$ is globally isometric via the matrix logarithm to the Euclidean space $(\mathrm{Sym}(m), \langle \cdot, \cdot \rangle_F)$, and hence $(\mathcal{S}_+^m, \langle \cdot, \cdot \rangle_\cdot^{\mathrm{LE}})$ is a Cartan-Hadamard manifold. Moreover, the sectional curvature of $(\mathcal{S}_+^m, \langle \cdot, \cdot \rangle_\cdot^{\mathrm{LE}})$ is constantly zero.

C Proof of Proposition 1

C.1 Congruence

After conjugation of the input matrix $\mathbf{X}$, the output of the multi-head BiMap layer is $\{\mathbf{W}_1^h \mathbf{U}\mathbf{X}\mathbf{U}^\top (\mathbf{W}_1^h)^\top : h = 1, \ldots, H\}$, where $\mathbf{U} \in \mathrm{O}(m)$ is an orthogonal matrix. Each weight matrix $\mathbf{W}_1^h$ is full row rank with orthonormal rows. Let $\tilde{\mathbf{X}}_2$ denote the Fréchet mean after congruence on the input $\mathbf{X}$. If there exists an orthogonal matrix $\mathbf{Q} \in \mathrm{O}(m_1)$ that $\mathbf{W}_1^h \mathbf{U} = \mathbf{Q}\mathbf{W}_1^h$ for all $1 \leq h \leq H$, then

$$\mathbf{W}_1^h \mathbf{U}\mathbf{X}\mathbf{U}^\top (\mathbf{W}_1^h)^\top = \mathbf{Q}\mathbf{W}_1^h \mathbf{X} (\mathbf{W}_1^h)^\top \mathbf{Q}^\top = \mathbf{Q}\mathbf{X}_1^h \mathbf{Q}^\top. \tag{12}$$

Under the LE Riemannian metric, we have $\log(\mathbf{Q}\mathbf{X}_1^h\mathbf{Q}^\top) = \mathbf{Q}\log(\mathbf{X}_1^h)\mathbf{Q}^\top$, and that the LE geodesic distance is invariant under orthogonal conjugation:

$$d_{\mathrm{LE}}(\mathbf{X}_2, \mathbf{X}_1^h) = d_{\mathrm{LE}}(\mathbf{Q}\mathbf{X}_2\mathbf{Q}^\top, \mathbf{Q}\mathbf{X}_1^h\mathbf{Q}^\top).$$

Therefore, after matrix congruence, the LE Fréchet mean is $\tilde{\mathbf{X}}_2 = \mathbf{Q}\mathbf{X}_2\mathbf{Q}^\top$.

Eq. (12) implies that conjugating the input by $\mathbf{U}$ is equivalent to conjugating every head's output by $\mathbf{Q}$. Let $\mathrm{Im}((\mathbf{W}_1^h)^\top) \subset \mathbb{R}^{m_1}$ denote the subspace spanned by the rows of $\mathbf{W}_1^h$. The relation $\mathbf{W}_1^h\mathbf{U} = \mathbf{Q}\mathbf{W}_1^h$ forces $\mathbf{U}$ $\mathrm{Im}((\mathbf{W}_1^h)^\top) = \mathrm{Im}((\mathbf{W}_1^h)^\top)$ and $\mathbf{W}_1^h\mathbf{U}(\mathbf{W}_1^h)^\top = \mathbf{Q}$; that is, each head's row-space is $\mathbf{U}$-invariant, and the "compressed" action of $\mathbf{U}$ onto that subspace is the same $\mathbf{Q}$ for all heads. With one head, such a $\mathbf{Q}$ always exists. With multiple heads, asking for the same $\mathbf{Q}$ across all h is strong: either the subspaces $\mathrm{Im}((\mathbf{W}_1^h)^\top)$ are all the same and $\mathbf{U}$ acts on that subspace as $\mathbf{Q}$, or $\mathbf{U}$ acts trivially on every $\mathrm{Im}((\mathbf{W}_1^h)^\top)$.

If each head rotates by its own $\mathbf{Q}_h \in \mathrm{O}(m_1)$: $\mathbf{W}_1^h\mathbf{U} = \mathbf{Q}_h\mathbf{W}_1^h$, assume additionally that the candidate Fréchet mean $\tilde{\mathbf{X}}_2$ commutes with every $\mathbf{Q}_h$: $\tilde{\mathbf{X}}_2\mathbf{Q}_h = \mathbf{Q}_h\tilde{\mathbf{X}}_2$. Then under the LE metric, we have

$$\begin{aligned}\sum_{h=1}^{H} \|\log(\tilde{\mathbf{X}}_2) - \log(\mathbf{Q}_h\mathbf{X}_1^h\mathbf{Q}_h^\top)\|_F^2 &= \sum_{h=1}^{H} \|\log(\tilde{\mathbf{X}}_2) - \mathbf{Q}_h\log(\mathbf{X}_1^h)\mathbf{Q}_h^\top\|_F^2 \\ &= \sum_{h=1}^{H} \|\log(\tilde{\mathbf{X}}_2) - \log(\mathbf{X}_1^h)\|_F^2.\end{aligned}$$

Therefore, the LE Fréchet mean is $\tilde{\mathbf{X}}_2 = \mathbf{X}_2$, invariant under general congruence.

The commutation requirement $\tilde{\mathbf{X}}_2\mathbf{Q}_h = \mathbf{Q}_h\tilde{\mathbf{X}}_2$ implies that $\tilde{\mathbf{X}}_2$ is invariant under all the head-wise rotations: $\tilde{\mathbf{X}}_2 = \mathbf{Q}_h\tilde{\mathbf{X}}_2\mathbf{Q}_h^\top$, for every h; that is, $\tilde{\mathbf{X}}_2$ shares a common orthonormal eigenbasis with all $\mathbf{Q}_h$ up to blocks: the space splits into invariant blocks for the group $\langle\mathbf{Q}_h\rangle$, and on each block $\tilde{\mathbf{X}}_2$ is a scalar multiple of the identity. In group-theory terms, $\tilde{\mathbf{X}}_2$ lies in the centralizer of the subgroup generated by $\{\mathbf{Q}_h : 1 \le h \le H\}$.

For the LC Riemannian metric, because the LC geometry depends on Cholesky factors of the matrices $\mathbf{X}_1^h = \mathbf{W}_1^h\mathbf{X}(\mathbf{W}_1^h)^\top$ (not on $\mathbf{X}$ directly), the inner congruence by $\mathbf{U} \in \mathrm{GL}(m)$ before compression does not translate to a uniform congruence on the $\mathbf{X}_1^h$. In particular, the LC Fréchet mean is defined by mapping each SPD $\mathbf{X}_1^h$ to its Cholesky factor $\lfloor\mathfrak{L}(\mathbf{X}_1^h)\rfloor + \log(\mathrm{diag}(\mathfrak{L}(\mathbf{X}_1^h)))$, averaging in that triangular space, then mapping back. This construction is not affine-invariant and not even orthogonally equivariant in general. Therefore, the LC Fréchet mean changes in a way that is not a simple transform of $\mathbf{X}_2$.

The affine-invariant (AI) Fréchet mean has congruence equivariance:

$$\mathbb{E}_{\mathrm{AI}}(\mathbf{Q}\mathbf{X}_1\mathbf{Q}^\top, \ldots, \mathbf{Q}\mathbf{X}_n\mathbf{Q}^\top) = \mathbf{Q}\mathbb{E}_{\mathrm{AI}}(\mathbf{X}_1, \ldots, \mathbf{X}_n)\mathbf{Q}^\top.$$

Therefore, the AI Fréchet mean is equivariant under matrix congruence, only if there exists a single $\mathbf{Q} \in \mathrm{GL}(m_1)$ that $\mathbf{W}_1^h\mathbf{U} = \mathbf{Q}\mathbf{W}_1^h$ for all $1 \le h \le H$.

C.2 Isometry

A diffeomorphism $\psi : \mathcal{S}_+^m \to \mathcal{S}_+^m$ is an isometry if $d_g(\psi(\mathbf{X}), \psi(\mathbf{Z})) = d_g(\mathbf{X}, \mathbf{Z})$ for all $\mathbf{X}, \mathbf{Z} \in \mathcal{S}_+^m$. Here, d_g is the geodesic distance induced by the LC or LE Riemannian metric. The isometries of $\mathcal{S}_+^m$ form a group under composition, denoted by $I(\mathcal{S}_+^m)$. Let $F : \mathcal{S}_+^m \to \mathcal{S}_+^{m_1}$ be the function defined by $\mathbf{X} \mapsto \mathbb{E}_g(\mathbf{W}_1^1\mathbf{X}(\mathbf{W}_1^1)^\top, \ldots, \mathbf{W}_1^H\mathbf{X}(\mathbf{W}_1^H)^\top)$, where $\mathbb{E}_g$ is either the LC or the LE Fréchet mean.

Given any two points $\mathbf{X}, \mathbf{Z} \in \mathcal{S}_+^m$, the LE geodesic between $F(\mathbf{X})$ and $F(\mathbf{Z})$ is

$$\begin{aligned}d_{\mathrm{LE}}(F(\mathbf{X}), F(\mathbf{Z})) &= \|\log(F(\mathbf{X})) - \log(F(\mathbf{Z}))\|_F \\ &= \|\frac{1}{H}\sum_{h=1}^{H}\log(\mathbf{W}_1^h\mathbf{X}(\mathbf{W}_1^h)^\top) - \frac{1}{H}\sum_{h=1}^{H}\log(\mathbf{W}_1^h\mathbf{Z}(\mathbf{W}_1^h)^\top)\|_F,\end{aligned}$$

and the LE geodesic between $F(\psi(\mathbf{X}))$ and $F(\psi(\mathbf{Z}))$ is

$$d_{\mathrm{LE}}(F(\psi(\mathbf{X})), F(\psi(\mathbf{Z}))) = \|\frac{1}{H}\sum_{h=1}^{H}\log(\mathbf{W}_1^h\psi(\mathbf{X})(\mathbf{W}_1^h)^\top) - \frac{1}{H}\sum_{h=1}^{H}\log(\mathbf{W}_1^h\psi(\mathbf{Z})(\mathbf{W}_1^h)^\top)\|_F,$$

Any LE-isometry is exactly a Euclidean isometry in the log domain: $\psi(\mathbf{X}) = \exp(\mathcal{O}(\log(\mathbf{X})) + \mathbf{U})$, where $\mathcal{O} : \mathrm{Sym}(m) \to \mathrm{Sym}(m)$ is linear orthogonal w.r.t. the Frobenius inner product, and $\mathbf{U} \in \mathrm{Sym}(m)$ is a fixed symmetric matrix (a translation in the log domain). The crucial observation is that for semi-orthogonal $\mathbf{W}_1^h$, $\log(\mathbf{W}_1^h \psi(\mathbf{X})(\mathbf{W}_1^h)^\top) \neq \mathbf{W}_1^h \log(\psi(\mathbf{X}))(\mathbf{W}_1^h)^\top$ in general. Hence the orthogonal action $\mathcal{O}$ and translation $\mathbf{U}$ in the log domain do not factor out through the $\mathbf{W}_1^h(\cdot)(\mathbf{W}_1^h)^\top$ compression, so the two distances above generally differ. Therefore, the map F is not invariant (nor equivariant) to the action of the LE-isometry group.

Likewise, any LC-isometry is exactly a Euclidean isometry in the LC chart domain: $\psi(\mathbf{X}) = \Phi^{-1}(\mathbf{R}\Phi(\mathbf{X}) + \mathbf{u})$, where $\mathbf{R} \in O(m(m+1)/2)$, $\mathbf{u} \in \mathbb{R}^{m(m+1)/2}$, and Φ is the log-Cholesky coordinate map for the LC metric. In general, $\mathfrak{L}(\mathbf{W}_1^h \psi(\mathbf{X})(\mathbf{W}_1^h)^\top) \neq \mathbf{W}_1^h \mathfrak{L}(\psi(\mathbf{X}))$, even for orthogonal $\mathbf{W}_1^h$. In fact, the Cholesky factor $\mathfrak{L}(\mathbf{W}_1^h \psi(\mathbf{X})(\mathbf{W}_1^h)^\top)$ depends nonlinearly on both $\psi(\mathbf{X})$ and $\mathbf{W}_1^h$ via a re-triangularization step. Consequently, the linear orthogonal action $\mathbf{R}$ and translation $\mathbf{u}$ in the LC chart do not commute with the compression $\mathbf{W}_1^h \psi(\mathbf{X})(\mathbf{W}_1^h)^\top$. Therefore, F is not invariant (nor equivariant) under the action of the LC-isometry group.

D Proof of Propositions 2-4

D.1 Proof of Proposition 2

Under the LE metric, we have $F(\mathbf{X}) = \exp(\frac{1}{H}\sum_{h=1}^H \log(\mathbf{W}_1^h \mathbf{X}(\mathbf{W}_1^h)^\top))$, and

$$\begin{aligned} d_{\mathrm{LE}}(F(\mathbf{X}), F(\mathbf{Z})) &= \| \frac{1}{H} \sum_{h=1}^H \left[\log(\mathbf{W}_1^h \mathbf{X}(\mathbf{W}_1^h)^\top) - \log(\mathbf{W}_1^h \mathbf{Z}(\mathbf{W}_1^h)^\top) \right] \|_F \\ &\leq \frac{1}{H} \sum_{h=1}^H \| \log(\mathbf{W}_1^h \mathbf{X}(\mathbf{W}_1^h)^\top) - \log(\mathbf{W}_1^h \mathbf{Z}(\mathbf{W}_1^h)^\top) \|_F. \end{aligned}$$

Let $\gamma(t) = \exp((1-t)\log(\mathbf{Z}) + t\log(\mathbf{X}))$ be the LE geodesic from $\mathbf{Z}$ to $\mathbf{X}$, and define $f_h(t) = \log(\mathbf{W}_1^h \gamma(t)(\mathbf{W}_1^h)^\top)$. The function $f_h(t)$ is absolutely continuous, and therefore we have

$$\| \log(\mathbf{W}_1^h \mathbf{X}(\mathbf{W}_1^h)^\top) - \log(\mathbf{W}_1^h \mathbf{Z}(\mathbf{W}_1^h)^\top) \|_F = \| f_h(1) - f_h(0) \|_F = \| \int_0^1 f_h'(t) dt \|_F \leq \int_0^1 \| f_h'(t) \|_F dt.$$

We below prove that

$$\| f_h'(t) \|_F \leq \kappa(\gamma(t)) \| \log(\mathbf{X}) - \log(\mathbf{Z}) \|_F = \kappa(\gamma(t)) d_{\mathrm{LE}}(\mathbf{X}, \mathbf{Z}), \tag{13}$$

where $\kappa(\gamma(t)) = \lambda_{\max}(\gamma(t))/\lambda_{\min}(\gamma(t))$ is the condition number of $\gamma(t)$. When $\mathbf{X}, \mathbf{Z} \in \mathcal{X}_{\alpha,\beta}$, we have $\gamma(t) \in \mathcal{X}_{\alpha,\beta}$ and hence $\kappa(\gamma(t)) \leq \beta/\alpha$. Then it directly follows from Eq. (13) that F is Lipschitz on $\mathcal{X}_{\alpha,\beta}$ with constant $L \leq \beta/\alpha$.

Write $\mathbf{U}(t) = \log(\gamma(t)) = (1-t)\log(\mathbf{Z}) + t\log(\mathbf{X})$ and $\mathbf{V} = \log(\mathbf{X}) - \log(\mathbf{Z})$. We have $\gamma'(t) = (D_{\mathbf{U}(t)} \exp)(\mathbf{V})$, and the chain rule gives

$$f_h'(t) = (D_{\mathbf{W}_1^h \gamma(t) (\mathbf{W}_1^h)^\top} \log) \left(\mathbf{W}_1^h (D_{\mathbf{U}(t)} \exp)(\mathbf{V}) (\mathbf{W}_1^h)^\top \right).$$

The differential $D_{\mathbf{X}} \log : \mathcal{T}_{\mathbf{X}} \mathcal{S}_+^m \mapsto \mathcal{T}_{\log(\mathbf{X})} \mathrm{Sym}(m)$ of the matrix logarithm at $\mathbf{X}$ has the integral form

$$(D_{\mathbf{X}} \log)(\mathbf{K}) = \int_0^\infty (\mathbf{X} + s\mathbf{I})^{-1} \mathbf{K} (\mathbf{X} + s\mathbf{I})^{-1} ds, \ \ \mathbf{K} \in \mathcal{T}_{\mathbf{X}} \mathcal{S}_+^m,$$

Taking Frobenius norms and using $\| (\mathbf{X} + s\mathbf{I})^{-1} \|_2 = \frac{1}{\lambda_{\min}(\mathbf{X}) + s}$, we have

$$\| (D_{\mathbf{X}} \log)(\mathbf{K}) \|_F \leq \int_0^\infty \| (\mathbf{X} + s\mathbf{I})^{-1} \|_2^2 \| \mathbf{K} \|_F ds = \| \mathbf{K} \|_F \int_0^\infty \frac{ds}{(\lambda_{\min}(\mathbf{X}) + s)^2} = \frac{\| \mathbf{K} \|_F}{\lambda_{\min}(\mathbf{X})}.$$

Therefore, we have

$$\| f_h'(t) \|_F \leq \frac{\| \mathbf{W}_1^h (D_{\mathbf{U}(t)} \exp)(\mathbf{V}) (\mathbf{W}_1^h)^\top \|_F}{\lambda_{\min}(\mathbf{W}_1^h \gamma(t) (\mathbf{W}_1^h)^\top)} \leq \frac{\| (D_{\mathbf{U}(t)} \exp)(\mathbf{V}) \|_F}{\lambda_{\min}(\mathbf{W}_1^h \gamma(t) (\mathbf{W}_1^h)^\top)},$$

where we have utilized the property that $\|\mathbf{XYZ}\|_F \le \|\mathbf{X}\|_2\|\mathbf{Y}\|_F\|\mathbf{Z}\|_2$ for any matrices $\{\mathbf{X}, \mathbf{Y}, \mathbf{Z}\}$, and $\|\mathbf{X}\|_2 = 1$ when $\mathbf{X}$ has orthonormal rows. Moreover, because $\mathbf{W}_1^h\gamma(t)(\mathbf{W}_1^h)^\top$ is a compression of $\gamma(t)$ by an isometry, eigenvalues interlace (Fan and Pall, 1957): $\lambda_{\min}(\mathbf{W}_1^h\gamma(t)(\mathbf{W}_1^h)^\top) \ge \lambda_{\min}(\gamma(t))$.

The differential $D_{\mathbf{U}}\exp : \mathcal{T}_{\mathbf{U}}\mathrm{Sym}(m) \mapsto \mathcal{T}_{\exp(\mathbf{U})}\mathcal{S}_+^m$ of matrix exponential map at $\mathbf{U}$ has the integral form

$$(D_{\mathbf{U}}\exp)(\mathbf{V}) = \int_0^1 \exp((1-s)\mathbf{U})\mathbf{V}\exp(s\mathbf{U})ds, \;\; \mathbf{V} \in \mathcal{T}_{\mathbf{X}}\mathcal{S}_+^m.$$

Taking Frobenius norms, we have

$$\|(D_{\mathbf{U}}\exp)(\mathbf{V})\|_F \le \int_0^1 \|\exp((1-s)\mathbf{U})\|_2\|\mathbf{V}\|_F\|\exp(s\mathbf{U})\|_2 ds = \|\mathbf{V}\|_F\|\exp(\mathbf{U})\|_2.$$

When $\mathbf{U} = \mathbf{U}(t)$, we have $\|\exp(\mathbf{U}(t))\|_2 = \|\gamma(t)\|_2 = \lambda_{\max}(\gamma(t))$, and substituting $\mathbf{V}$ with $\log(\mathbf{X}) - \log(\mathbf{Z})$, we have

$$\|f_h'(t)\|_F \le \frac{\lambda_{\max}(\gamma(t))\|\log(\mathbf{X}) - \log(\mathbf{Z})\|_F}{\lambda_{\min}(\mathbf{W}_1^h\gamma(t)(\mathbf{W}_1^h)^\top)} \le \frac{\lambda_{\max}(\gamma(t))}{\lambda_{\min}(\gamma(t))}\|\log(\mathbf{X}) - \log(\mathbf{Z})\|_F.$$

D.2 Proof of Proposition 3

The log-Cholesky coordinate map $\Phi : \mathcal{S}_+^m \to \mathbb{R}^{m(m+1)/2}$, $\Phi(\mathbf{X}) = (\mathrm{vnz}(\lfloor\mathfrak{L}(\mathbf{X})\rfloor), \log(\mathrm{vd}(\mathfrak{L}(\mathbf{X}))))$ is a global coordinate map, where $\mathrm{vnz}(\lfloor\mathfrak{L}(\mathbf{X})\rfloor)$ is the vector of the non-zero entries of $\lfloor\mathfrak{L}(\mathbf{X})\rfloor$, and $\mathrm{vd}(\mathfrak{L}(\mathbf{X}))$ is the vector of the diagonal entries of $\mathfrak{L}(\mathbf{X})$. The LC metric makes $\mathcal{S}_+^m$ flat via the chart map Φ, and the LC geodesic distance is simply the Euclidean norm in the log-Cholesky coordinate: $d_{\mathrm{LC}}(\mathbf{X}, \mathbf{Z}) = \|\Phi(\mathbf{X}) - \Phi(\mathbf{Z})\|_2$. For any $\mathbf{X}, \mathbf{Z} \in \mathcal{S}_+^m$, we have

$$\begin{aligned} d_{\mathrm{LC}}(F(\mathbf{X}), F(\mathbf{Z})) &= \|\frac{1}{H}\sum_{h=1}^H \left[\Phi(\mathbf{W}_1^h\mathbf{X}(\mathbf{W}_1^h)^\top) - \Phi(\mathbf{W}_1^h\mathbf{Z}(\mathbf{W}_1^h)^\top)\right]\|_2 \\ &\le \frac{1}{H}\sum_{h=1}^H \|\Phi(\mathbf{W}_1^h\mathbf{X}(\mathbf{W}_1^h)^\top) - \Phi(\mathbf{W}_1^h\mathbf{Z}(\mathbf{W}_1^h)^\top)\|_2. \end{aligned}$$

In the proof below, we drop the indices and write $\mathbf{W}$ for $\mathbf{W}_1^h$. We have $\mathbf{W}\mathbf{X}\mathbf{W}^\top = (\mathbf{W}\mathfrak{L}(\mathbf{X}))(\mathbf{W}\mathfrak{L}(\mathbf{X}))^\top$, and a thin QR decomposition: $(\mathbf{W}\mathfrak{L}(\mathbf{X}))^\top = \mathbf{Q}_{\mathbf{X}}\mathbf{R}_{\mathbf{X}}$, where $\mathbf{Q}_{\mathbf{X}} \in \mathbb{R}^{m\times m_1}$ has orthonormal columns, and $\mathbf{R}_{\mathbf{X}} \in \mathbb{R}^{m_1\times m_1}$ is upper-triangular with positive diagonal. The Cholesky factor of $\mathbf{W}\mathbf{X}\mathbf{W}^\top$ is exactly $\mathbf{R}_{\mathbf{X}}^\top$: $\mathfrak{L}(\mathbf{W}\mathbf{X}\mathbf{W}^\top) = \mathbf{R}_{\mathbf{X}}^\top$. Therefore, we have

$$\|\Phi(\mathbf{W}\mathbf{X}\mathbf{W}^\top) - \Phi(\mathbf{W}\mathbf{Z}\mathbf{W}^\top)\|_2^2 = \|\lfloor\mathbf{R}_{\mathbf{X}}^\top\rfloor - \lfloor\mathbf{R}_{\mathbf{Z}}^\top\rfloor\|_F^2 + \|\log(\mathrm{diag}(\mathbf{R}_{\mathbf{X}}^\top)) - \log(\mathrm{diag}(\mathbf{R}_{\mathbf{Z}}^\top))\|_F^2.$$

Fix any spectral band $0 < \alpha \le \beta < \infty$, and consider the subset $\mathcal{X}_{\alpha,\beta} = \{\mathbf{X} \in \mathcal{S}_+^m : \alpha\mathbf{I} \preceq \mathbf{X} \preceq \beta\mathbf{I}\}$. Then we have

$$\begin{aligned} \|\Phi(\mathbf{W}\mathbf{X}\mathbf{W}^\top) - \Phi(\mathbf{W}\mathbf{Z}\mathbf{W}^\top)\|_2^2 &\le \|\lfloor\mathbf{R}_{\mathbf{X}}^\top\rfloor - \lfloor\mathbf{R}_{\mathbf{Z}}^\top\rfloor\|_F^2 + \alpha^{-1}\|\mathrm{diag}(\mathbf{R}_{\mathbf{X}}^\top) - \mathrm{diag}(\mathbf{R}_{\mathbf{Z}}^\top)\|_F^2 \\ &\le \max\{1, \alpha^{-1}\}\|\mathbf{R}_{\mathbf{X}}^\top - \mathbf{R}_{\mathbf{Z}}^\top\|_F^2. \end{aligned}$$

Define a line in the $m \times m$ matrix space: $\mathbf{Y}(t) = (1-t)\mathbf{Z} + t\mathbf{X}$, for $0 \le t \le 1$, and let $\mathbf{A}(t) = \mathbf{W}\mathbf{Y}(t)\mathbf{W}^\top$. It can be readily proved that $\mathbf{A}(t) \in \mathcal{X}_{\alpha,\beta}$, for any $0 \le t \le 1$. The derivative in t is $\dot{\mathbf{A}}(t) = \mathbf{W}(\mathbf{X} - \mathbf{Z})\mathbf{W}^\top$.

Let $\mathbf{L}(t)$ be the Cholesky factor of $\mathbf{A}(t)$: $\mathfrak{L}(\mathbf{A}(t)) = \mathbf{L}(t)$. In particular, $\mathbf{L}(0) = \mathbf{R}_{\mathbf{Z}}^\top$ and $\mathbf{L}(1) = \mathbf{R}_{\mathbf{X}}^\top$. Differentiate the Cholesky map $\mathbf{A}(t) = \mathbf{L}(t)\mathbf{L}(t)^\top$: $\dot{\mathbf{A}}(t) = \dot{\mathbf{L}}(t)\mathbf{L}(t)^\top + \mathbf{L}(t)\dot{\mathbf{L}}(t)^\top$, and then we have

$$\mathbf{L}(t)^{-1}\dot{\mathbf{A}}(t)\mathbf{L}(t)^{-\top} = \mathbf{L}(t)^{-1}\dot{\mathbf{L}}(t) + \dot{\mathbf{L}}(t)^\top\mathbf{L}(t)^{-\top} = \mathbf{\Omega} + \mathbf{\Omega}^\top,$$

where $\mathbf{\Omega} = \mathbf{L}(t)^{-1}\dot{\mathbf{L}}(t)$ is lower-triangular. Therefore, we have $\|\mathbf{\Omega}\|_F \le \|\mathbf{\Omega} + \mathbf{\Omega}^\top\|_F \le \|\mathbf{L}(t)^{-1}\|_2^2\|\dot{\mathbf{A}}(t)\|_F$. Finally $\dot{\mathbf{L}}(t) = \mathbf{L}(t)\mathbf{\Omega}$ gives

$$\|\dot{\mathbf{L}}(t)\|_F \le \|\mathbf{L}(t)\|_2\|\mathbf{\Omega}\|_F \le \|\mathbf{L}(t)\|_2\|\mathbf{L}(t)^{-1}\|_2^2\|\dot{\mathbf{A}}(t)\|_F.$$

The squared singular values of $\mathbf{L}(t)$ are the eigen values of $\mathbf{A}(t)$, and therefore $\|\mathbf{L}(t)\|_2 \leq \sqrt{\beta}$ and $\|\mathbf{L}(t)^{-1}\|_2 \leq \frac{1}{\sqrt{\alpha}}$. Finally, we have $\|\dot{\mathbf{L}}(t)\|_F \leq \frac{\sqrt{\beta}}{\alpha}\|\dot{\mathbf{A}}(t)\|_F$.

Integrating along the path, we have

$$\|\mathbf{R}_{\mathbf{X}}^\top - \mathbf{R}_{\mathbf{Z}}^\top\|_F \leq \int_0^1 \|\dot{\mathbf{L}}(t)\|_F dt \leq \frac{\sqrt{\beta}}{\alpha}\int_0^1 \|\dot{\mathbf{A}}(t)\|_F dt = \frac{\sqrt{\beta}}{\alpha}\|\mathbf{W}(\mathbf{X}-\mathbf{Z})\mathbf{W}^\top\|_F.$$

Given that $\mathbf{W}\mathbf{W}^\top = \mathbf{I}$, we have $\|\mathbf{W}(\mathbf{X}-\mathbf{Z})\mathbf{W}^\top\|_F \leq \|\mathbf{W}\|_2\|\mathbf{X}-\mathbf{Z}\|_F\|\mathbf{W}^\top\|_2 = \|\mathbf{X}-\mathbf{Z}\|_F$. Write $\mathbf{X}-\mathbf{Z} = \mathfrak{L}(\mathbf{X})(\mathfrak{L}(\mathbf{X})-\mathfrak{L}(\mathbf{Z}))^\top + (\mathfrak{L}(\mathbf{X})-\mathfrak{L}(\mathbf{Z}))\mathfrak{L}(\mathbf{Z})^\top$, we have

$$\|\mathbf{X}-\mathbf{Z}\|_F \leq (\|\mathfrak{L}(\mathbf{X})\|_2 + \|\mathfrak{L}(\mathbf{Z})\|_2)\|\mathfrak{L}(\mathbf{X})-\mathfrak{L}(\mathbf{Z})\|_F \leq 2\sqrt{\beta}\|\mathfrak{L}(\mathbf{X})-\mathfrak{L}(\mathbf{Z})\|_F.$$

Applying the mean-value theorem on the diagonal entries, we have

$$\begin{aligned}|[\mathfrak{L}(\mathbf{X})]_{i,i} - [\mathfrak{L}(\mathbf{Z})]_{i,i}| &\leq \max\{[\mathfrak{L}(\mathbf{X})]_{i,i}, [\mathfrak{L}(\mathbf{Z})]_{i,i}\}|\log([\mathfrak{L}(\mathbf{X})]_{i,i}) - \log([\mathfrak{L}(\mathbf{Z})]_{i,i})| \\ &\leq \sqrt{\beta}|\log([\mathfrak{L}(\mathbf{X})]_{i,i}) - \log([\mathfrak{L}(\mathbf{Z})]_{i,i})|,\end{aligned}$$

and therefore

$$\begin{aligned}\|\mathfrak{L}(\mathbf{X})-\mathfrak{L}(\mathbf{Z})\|_F^2 &\leq \|\lfloor\mathfrak{L}(\mathbf{X})\rfloor - \lfloor\mathfrak{L}(\mathbf{Z})\rfloor\|_F^2 + \beta\|\log(\mathrm{diag}(\mathfrak{L}(\mathbf{X}))) - \log(\mathrm{diag}(\mathfrak{L}(\mathbf{Z})))\|_F^2 \\ &\leq \beta\|\Phi(\mathbf{X})-\Phi(\mathbf{Z})\|_2^2\end{aligned}$$

Putting everything together, we have

$$\|\Phi(\mathbf{W}\mathbf{X}\mathbf{W}^\top) - \Phi(\mathbf{W}\mathbf{Z}\mathbf{W}^\top)\|_2 \leq \max\{1, \alpha^{-1}\}\frac{2\beta^{3/2}}{\alpha}\|\Phi(\mathbf{X})-\Phi(\mathbf{Z})\|_2,$$

for any $\mathbf{W}$ of orthonormal rows.

D.3 Proof of Proposition 4

Let ϕ be either the matrix logarithm map $\log$ or the log-Cholesky coordinate map Φ. The unified formula for the Fréchet mean in the Pooling layer is

$$F(\mathbf{X}; \{\mathbf{W}_1^h\}_{h=1}^H) = \phi^{-1}(\frac{1}{H}\sum_{h=1}^H \phi(\mathbf{W}_1^h\mathbf{X}(\mathbf{W}_1^h)^\top)).$$

Write $\mathbf{Z} = F(\mathbf{X}; \{\mathbf{W}_1^h\}_{h=1}^H)$. If the function F is "linear", then there exists another weight matrix $\mathbf{W} \in \mathbb{R}^{m_1\times m}$ with $\mathbf{W}\mathbf{W}^\top = \mathbf{I}$ such that $\mathbf{Z} = F(\mathbf{X}; \{\mathbf{W}_1^h\}_{h=1}^H) = \mathbf{W}\mathbf{X}\mathbf{W}^\top$.

Let $\lambda_1(\mathbf{X}) \geq ... \geq \lambda_m(\mathbf{X})$ be the eigenvalues of $\mathbf{X}$, and $\mu_1(\mathbf{Z}) \geq ... \geq \mu_{m_1}(\mathbf{Z})$ be the eigenvalues of $\mathbf{Z}$. According to Fan and Pall (1957), if and only if the eigenvalues of $\mathbf{Z}$ satisfy the interlacing inequalities: $\lambda_j(\mathbf{X}) \geq \mu_j(\mathbf{Z}) \geq \lambda_{j+(m-m_1)}(\mathbf{X})$, for any $j = 1, \ldots, m_1$, then there exists an $\mathbf{W} \in \mathbb{R}^{m_1\times m}$ with $\mathbf{W}\mathbf{W}^\top = \mathbf{I}$ such that $\mathbf{Z} = \mathbf{W}\mathbf{X}\mathbf{W}^\top$.

Under the LE metric, with the eigenvalue decomposition $\mathbf{W}_1^h\mathbf{X}(\mathbf{W}_1^h)^\top = \mathbf{Q}_1^h\mathbf{\Lambda}_1^h(\mathbf{Q}_1^h)^\top$, the Fréchet mean is

$$\mathbf{Z} = \exp(\frac{1}{H}\sum_{j=1}^H \mathbf{Q}_1^h\log(\mathbf{\Lambda}_1^h)(\mathbf{Q}_1^h)^\top).$$

Note that, in general, the heads $\{\mathbf{W}_1^h\mathbf{X}(\mathbf{W}_1^h)^\top : h = 1, \ldots, H\}$ do not have a common orthonormal eigenbasis. Therefore, although each compression $\mathbf{W}_1^h\mathbf{X}(\mathbf{W}_1^h)^\top$ individually interlaces with $\mathbf{X}$, the eigenvalues of the average $\mathbf{Z}$ in general do not satisfy the interlacing inequalities.

The key point is that $\Omega := \{\log(\mathbf{W}\mathbf{X}\mathbf{W}^\top) : \mathbf{W}\mathbf{W}^\top = \mathbf{I}\}$ is not a convex set for $m_1 \geq 2$, and hence taking average of $\{\log(\mathbf{W}_1^h\mathbf{X}(\mathbf{W}_1^h)^\top) : h = 1, \ldots, H\}$ can leave the set Ω. Therefore, in general there is no $\mathbf{W}$ with $\mathbf{W}\mathbf{W}^\top = \mathbf{I}$ such that $\log(\mathbf{W}\mathbf{X}\mathbf{W}^\top) = \frac{1}{H}\sum_{h=1}^H \log(\mathbf{W}_1^h\mathbf{X}(\mathbf{W}_1^h)^\top))$. Since $\exp$ and $\log$ are inverse diffeomorphisms between $\mathcal{S}_+^{m_1}$ and $\mathrm{Sym}(m_1)$, this implies there is no such $\mathbf{W}$ with $F(\mathbf{X}; \{\mathbf{W}_1^h\}_{h=1}^H) = \mathbf{W}\mathbf{X}\mathbf{W}^\top$. Consequently, the spectrum of the average $\frac{1}{H}\sum_{j=1}^H \mathbf{Q}_1^h\log(\mathbf{\Lambda}_1^h)(\mathbf{Q}_1^h)^\top$ need not satisfy the Fan-Pall interlacing inequalities relative to $\mathbf{X}$.

An identical argument holds for the LC metric by replacing $\log$ with Φ and noting that the set $\{\Phi(\mathbf{W}\mathbf{X}\mathbf{W}^\top) : \mathbf{W}\mathbf{W}^\top = \mathbf{I}\}$ is likewise non-convex.

E BACKPROPAGATION IN THE MATRIX-TO-VECTOR MODULE

We absorb each BiMap layer into its succeeding operation, yielding two composite maps $f^{ap}(\cdot)$ and $f^{log}(\cdot)$. In particular, $f^{ap}(\cdot)$ directly maps $\mathbf{X}(t)$ to $\mathbf{X}_3(t)$:

$$\mathbf{X}_3(t) = f^{ap}(\mathbf{X}(t); \{\mathbf{W}_1^h\}_{h=1}^H) = \phi^{-1}(\frac{\alpha}{H}\sum_{h=1}^{H}\phi(\mathbf{W}_1^h\mathbf{X}(t)(\mathbf{W}_1^h)^\top)),$$

then $f^{log}(\cdot)$ directly maps $\mathbf{X}_3(t)$ to $\mathbf{Y}^h(t)$:

$$\mathbf{Y}^h(t) = f^{log}(\mathbf{X}_3(t); \mathbf{W}_4^h) = \mathrm{Log}_{\mathbf{I}}^g(\mathbf{W}_4^h\mathbf{X}_3(t)(\mathbf{W}_4^h)^\top), \;\; h = 1, \ldots, H.$$

Note that both $f^{ap}(\cdot)$ and $f^{log}(\cdot)$ act pointwise in time: they are applied independently at each t to the matrix $\mathbf{X}(t)$ or $\mathbf{X}_3(t)$, rather than being functionals of the entire trajectory $\mathbf{X} \in \mathcal{H}_g([t_0, t_1]; \mathcal{S}_+^m)$. Consequently, the derivatives with respect to the inputs, namely $\frac{\partial f^{ap}}{\partial \mathbf{X}(t)}$ and $\frac{\partial f^{log}}{\partial \mathbf{X}_3(t)}$, are the usual matrix Jacobians evaluated at time t (as in the static, single-matrix case), rather than functional/variational derivatives with respect to the path.

However, the functional layer in the encoder $\mathcal{E}$, namely Eq. 2, is a functional of the input vector-valued function. We write $\boldsymbol{z} = \int_{t_0}^{t_1} \boldsymbol{W}^{(1)}(t)\boldsymbol{y}(t)dt + \boldsymbol{b}$ and $\boldsymbol{x}^{(1)} = a(\boldsymbol{z})$. For a perturbation $\boldsymbol{y} \to \boldsymbol{y} + \epsilon\boldsymbol{\eta}$, the first variation is

$$\delta\boldsymbol{z} = \int_{t_0}^{t_1} \boldsymbol{W}^{(1)}(t)\boldsymbol{\eta}(t)dt, \quad \delta\boldsymbol{x}^{(1)} = \mathrm{diag}(a'(\boldsymbol{z}))\delta\boldsymbol{z} = \int_{t_0}^{t_1} [\mathrm{diag}(a'(\boldsymbol{z}))\boldsymbol{W}^{(1)}(t)]\boldsymbol{\eta}(t)dt,$$

where $\mathrm{diag}(a'(\boldsymbol{z}))$ is the Jacobian of the element-wise activation a at $\boldsymbol{z}$. Hence, the functional derivative density with respect to the whole function $\boldsymbol{y}$ is the matrix kernel $\frac{\partial \boldsymbol{x}^{(1)}}{\partial \boldsymbol{y}}(s) = \mathrm{diag}(a'(\boldsymbol{z}))\boldsymbol{W}^{(1)}(s) \in \mathbb{R}^{q_1 \times p}$ for $s \in [t_0, t_1]$, where q_1 is the dimension of $\boldsymbol{x}^{(1)}$.

Recall that $\boldsymbol{y}(t)^\top = [\mathrm{vl}(\mathbf{Y}^1(t))^\top, \ldots, \mathrm{vl}(\mathbf{Y}^H(t))^\top]$ and vl is an isometry. Now split the matrix kernel into H blocks: $\mathrm{diag}(a'(\boldsymbol{z}))\boldsymbol{W}^{(1)}(s) = [\mathbf{K}_1(s), \ldots, \mathbf{K}_H(s)]$, where $\mathbf{K}_h(s) \in \mathbb{R}^{q_1 \times m_2(m_2+1)/2}$ and $p = H \times m_2(m_2+1)/2$. Then the functional derivative density w.r.t. the whole $\mathbf{Y}^h$ is

$$\frac{\partial x_r^{(1)}}{\partial \mathbf{Y}^h}(s) = \mathrm{vl}^{-1}([\mathbf{K}_h(s)]_{r,:}^\top), \text{ for } r = 1, \ldots, q_1.$$

E.1 LOGARITHM LAYER

For each head $1 \le h \le H$, we have $\mathbf{Y}^h(t) = \mathrm{Log}_{\mathbf{I}}^g(\mathbf{W}_4^h\mathbf{X}_3(t)(\mathbf{W}_4^h)^\top)$, where $(\mathbf{W}_4^h)^\top \in \mathrm{St}(m_1, m_2)$. To simplify notation, we drop the head index h and write $\mathbf{X}_4(t) = \mathbf{W}_4\mathbf{X}_3(t)\mathbf{W}_4^\top$.

LE metric When g is the LE metric, the Riemannian logarithm at the identity is exactly the matrix logarithm: $\mathrm{Log}_{\mathbf{I}}^g(\mathbf{X}_4(t)) = \log(\mathbf{X}_4(t))$. The Fréchet derivative of log at $\mathbf{X}_4(t)$ applied to a direction $\mathbf{U}$ is

$$(D_{\mathbf{X}_4(t)}\log)(\mathbf{U}) = \int_0^\infty [\mathbf{X}_4(t) + s\mathbf{I}]^{-1}\mathbf{U}[\mathbf{X}_4(t) + s\mathbf{I}]^{-1}ds,$$

which is self-adjoint w.r.t. the Frobenius inner product. If $\mathbf{U}$ is symmetric (resp. skew), then $(D_{\mathbf{X}_4(t)}\log)(\mathbf{U})$ is symmetric (resp. skew). The spectral form for the Fréchet derivative $(D_{\mathbf{X}_4(t)}\log)(\mathbf{U})$ is given in Eq. 8. Then, by the chain rule $D_{\mathbf{W}_4}\mathbf{Y}(t) = D_{\mathbf{X}_4(t)}\log \circ D_{\mathbf{W}_4}\mathbf{X}_4(t)$, the directional derivative with respect to $\mathbf{W}_4$ in direction $\Delta\mathbf{W}$ is

$$(D_{\mathbf{W}_4}\mathbf{Y}(t))(\Delta\mathbf{W}) = (D_{\mathbf{X}_4(t)}\log)(\Delta\mathbf{W}\mathbf{X}_3(t)\mathbf{W}_4^\top + \mathbf{W}_4\mathbf{X}_3(t)\Delta\mathbf{W}^\top).$$

Let $\mathbf{V}(t)$ be an "upstream gradient", and write $f(t) = \langle \mathbf{V}(t), \mathbf{Y}(t)\rangle_F$. The linear mapping $(D_{\mathbf{X}_4(t)}\log)(\cdot)$ is self-adjoint, and therefore we have

$$\begin{aligned} df(t) = \langle \mathbf{V}(t), d\mathbf{Y}(t)\rangle_F &= \langle \mathbf{V}(t), (D_{\mathbf{X}_4(t)}\log)(d\mathbf{X}_4(t))\rangle_F \\ &= \langle \mathbf{V}_s(t), (D_{\mathbf{X}_4(t)}\log)(d\mathbf{X}_4(t))\rangle_F = \langle (D_{\mathbf{X}_4(t)}\log)(\mathbf{V}_s(t)), d\mathbf{X}_4(t)\rangle_F, \end{aligned}$$

where $d\mathbf{X}_4(t) = (d\mathbf{W}_4)\mathbf{X}_3(t)\mathbf{W}_4^\top + \mathbf{W}_4\mathbf{X}_3(t)(d\mathbf{W}_4)^\top$, and $\mathbf{V}_s(t) = \frac{1}{2}(\mathbf{V}(t) + \mathbf{V}(t)^\top)$. Then we obtain

$$\begin{aligned} df(t) &= \mathrm{tr}((D_{\mathbf{X}_4(t)}\log)(\mathbf{V}_s(t))(d\mathbf{W}_4)\mathbf{X}_3(t)\mathbf{W}_4^\top) + \mathrm{tr}((D_{\mathbf{X}_4(t)}\log)(\mathbf{V}_s(t))\mathbf{W}_4\mathbf{X}_3(t)(d\mathbf{W}_4)^\top) \\ &= \mathrm{tr}(\mathbf{X}_3(t)\mathbf{W}_4^\top(D_{\mathbf{X}_4(t)}\log)(\mathbf{V}_s(t))(d\mathbf{W}_4)) + \mathrm{tr}(((D_{\mathbf{X}_4(t)}\log)(\mathbf{V}_s(t))\mathbf{W}_4\mathbf{X}_3(t))^\top(d\mathbf{W}_4)) \\ &= \langle 2(D_{\mathbf{X}_4(t)}\log)(\mathbf{V}_s(t))\mathbf{W}_4\mathbf{X}_3(t), d\mathbf{W}_4\rangle_F. \end{aligned}$$

Then the **Euclidean** gradient w.r.t. $\mathbf{W}_4$ is

$$\nabla_{\mathbf{W}_4} f(t) = 2(D_{\mathbf{X}_4(t)}\log)(\mathbf{V}_s(t))\mathbf{W}_4\mathbf{X}_3(t). \tag{14}$$

Then we can project $\nabla_{\mathbf{W}_4} f(t)$ onto the tangent space to get the Riemannian gradient.

When taking directional derivative with respect to $\mathbf{X}_3(t)$, we have

$$\begin{aligned} df(t) = \langle \mathbf{V}(t), (D_{\mathbf{X}_4(t)}\log)(d\mathbf{X}_4(t))\rangle_F &= \langle (D_{\mathbf{X}_4(t)}\log)(\mathbf{V}_s(t)), d\mathbf{X}_4(t)\rangle_F \\ &= \langle (D_{\mathbf{X}_4(t)}\log)(\mathbf{V}_s(t)), \mathbf{W}_4 d\mathbf{X}_3(t)\mathbf{W}_4^\top\rangle_F \\ &= \langle \mathbf{W}_4^\top(D_{\mathbf{X}_4(t)}\log)(\mathbf{V}_s(t))\mathbf{W}_4, d\mathbf{X}_3(t)\rangle_F. \end{aligned}$$

The Euclidean gradient w.r.t. the input $\mathbf{X}_3(t)$ is

$$\nabla_{\mathbf{X}_3(t)} f(t) = \mathbf{W}_4^\top(D_{\mathbf{X}_4(t)}\log)(\mathbf{V}_s(t))\mathbf{W}_4. \tag{15}$$

LC metric Let $\mathbf{L}(t) = \mathcal{L}(\mathbf{X}_4(t))$ denote the Cholesky factor of $\mathbf{X}_4(t) = \mathbf{W}_4\mathbf{X}_3(t)\mathbf{W}_4^\top$. Then $\mathbf{Y}(t) = \mathrm{Log}^{\mathrm{LC}}_{\mathbf{I}}(\mathbf{X}_4(t)) = 2\log(\mathrm{diag}(\mathbf{L}(t))) + \lfloor \mathbf{L}(t)\rfloor + \lfloor \mathbf{L}(t)\rfloor^\top$, and hence

$$d\mathbf{Y}(t) = 2\mathrm{diag}(\frac{dL_{1,1}(t)}{L_{1,1}(t)}, \cdots, \frac{dL_{m_2,m_2}(t)}{L_{m_2,m_2}(t)}) + \lfloor d\mathbf{L}(t)\rfloor + \lfloor d\mathbf{L}(t)\rfloor^\top.$$

Let $\mathbf{V}(t)$ be an "upstream gradient", and write $f(t) = \langle \mathbf{V}(t), \mathbf{Y}(t)\rangle_F$. Then the differential is

$$\begin{aligned} df(t) = \langle \mathbf{V}(t), d\mathbf{Y}(t)\rangle_F &= 2\sum_{i=1}^{m_2} V_{i,i}(t)\frac{dL_{i,i}(t)}{L_{i,i}(t)} + \sum_{i>j}[V_{i,j}(t) + V_{j,i}(t)]dL_{i,j}(t) \\ &= \langle \mathbf{H}(t), d\mathbf{L}(t)\rangle_F, \end{aligned}$$

where $\mathbf{H}(t)$ is a lower-triangular matrix: $H_{i,i}(t) = \frac{2V_{i,i}(t)}{L_{i,i}(t)}$, $H_{i,j}(t) = V_{i,j}(t) + V_{j,i}(t)$ $(i > j)$, and $H_{i,j}(t) = 0$ $(i < j)$.

The differential of the Cholesky factor satisfies $d\mathbf{L}(t) = \mathbf{L}(t)\Psi(\mathbf{L}(t)^{-1}d\mathbf{X}_4(t)\mathbf{L}(t)^{-\top})$, where Ψ is the lower-symmetrizer (keep strictly lower part and take half of the diagonal; zero upper part). Since Ψ is self-adjoint, the adjoint mapping gives

$$\begin{aligned} \langle \mathbf{H}(t), d\mathbf{L}(t)\rangle_F &= \langle \mathbf{L}(t)^\top\mathbf{H}(t), \Psi(\mathbf{L}(t)^{-1}d\mathbf{X}_4(t)\mathbf{L}(t)^{-\top})\rangle_F \\ &= \langle \Psi(\mathbf{L}(t)^\top\mathbf{H}(t)), \mathbf{L}(t)^{-1}d\mathbf{X}_4(t)\mathbf{L}(t)^{-\top}\rangle_F \\ &= \langle \mathbf{L}(t)^{-\top}\Psi(\mathbf{L}(t)^\top\mathbf{H}(t))^\top\mathbf{L}(t)^{-1}, d\mathbf{X}_4(t)\rangle_F \end{aligned}$$

Because $d\mathbf{X}_4(t)$ is symmetric, we have $\langle \mathbf{H}(t), d\mathbf{L}(t)\rangle_F = \langle \mathbf{U}(t), d\mathbf{X}_4(t)\rangle_F$, where $\mathbf{U}(t) = \frac{1}{2}\mathbf{L}(t)^{-\top}(\Psi(\mathbf{L}(t)^\top\mathbf{H}(t)) + \Psi(\mathbf{L}(t)^\top\mathbf{H}(t))^\top)\mathbf{L}(t)^{-1}$.

When taking directional derivative with respect to $\mathbf{W}_4$, we have

$$\begin{aligned} df(t) = \langle \mathbf{U}(t), d\mathbf{X}_4(t)\rangle_F &= \langle \mathbf{U}(t), (d\mathbf{W}_4)\mathbf{X}_3(t)\mathbf{W}_4^\top + \mathbf{W}_4\mathbf{X}_3(t)(d\mathbf{W}_4)^\top\rangle_F \\ &= \langle 2\mathbf{U}(t)\mathbf{W}_4\mathbf{X}_3(t), d\mathbf{W}_4\rangle_F. \end{aligned}$$

Therefore, the **Euclidean** gradient w.r.t. $\mathbf{W}_4$ is

$$\nabla_{\mathbf{W}_4} f(t) = 2\mathbf{U}(t)\mathbf{W}_4\mathbf{X}_3(t). \tag{16}$$

When taking directional derivative with respect to $\mathbf{X}_3(t)$, we have

$$df(t) = \langle \mathbf{U}(t), d\mathbf{X}_4(t)\rangle_F = \langle \mathbf{U}(t), \mathbf{W}_4 d\mathbf{X}_3(t)\mathbf{W}_4^\top\rangle_F = \langle \mathbf{W}_4^\top\mathbf{U}(t)\mathbf{W}_4, d\mathbf{X}_3(t)\rangle_F.$$

Therefore, the Euclidean gradient w.r.t. $\mathbf{X}_3(t)$ is

$$\nabla_{\mathbf{X}_3(t)} f(t) = \mathbf{W}_4^\top\mathbf{U}(t)\mathbf{W}_4. \tag{17}$$

E.2 ACTIVATED-POOLING LAYER

For the geodesic-shrinkage activation, we have

$$\mathbf{X}_3(t) = \phi^{-1}(\frac{\alpha}{H}\sum_{h=1}^{H}\phi(\mathbf{W}_1^h\mathbf{X}(t)(\mathbf{W}_1^h)^\top)),$$

where ϕ is the principal matrix logarithm log for the LE metric, and the log-Cholesky coordinate map Φ for the LC metric. Let $\mathbf{A}(t) = \frac{\alpha}{H}\sum_{h=1}^{H}\phi(\mathbf{W}_1^h\mathbf{X}(t)(\mathbf{W}_1^h)^\top)$ and $\mathbf{X}_1^h(t) = \mathbf{W}_1^h\mathbf{X}(t)(\mathbf{W}_1^h)^\top$. By the chain rule, for any h, we have

$$(D_{\mathbf{W}_1^h}\mathbf{X}_3(t))(\Delta\mathbf{W}) = (D_{\mathbf{A}(t)}\phi^{-1})(\frac{\alpha}{H}(D_{\mathbf{X}_1^h(t)}\phi)(\Delta\mathbf{W}\mathbf{X}(t)(\mathbf{W}_1^h)^\top + \mathbf{W}_1^h\mathbf{X}(t)\Delta\mathbf{W}^\top)).$$

LE metric Under the LE metric, we have

$$(D_{\mathbf{W}_1^h}\mathbf{X}_3(t))(\Delta\mathbf{W}) = (D_{\mathbf{A}(t)}\exp)(\frac{\alpha}{H}(D_{\mathbf{X}_1^h(t)}\log)(\Delta\mathbf{W}\mathbf{X}(t)(\mathbf{W}_1^h)^\top + \mathbf{W}_1^h\mathbf{X}(t)\Delta\mathbf{W}^\top)).$$

The spectral form for the Fréchet derivative $(D_{\mathbf{X}}\log)(\mathbf{U})$ is given in Eq. 8, and for $(D_{\mathbf{U}}\exp)(\mathbf{K})$ given in Eq. 9.

Let $\mathbf{V}(t)$ be an "upstream gradient", and write $f(t) = \langle\mathbf{V}(t), \mathbf{X}_3(t)\rangle_F$. The mapping $(D_{\mathbf{U}}\exp)(\cdot)$ is self-adjoint, and therefore we have

$$df(t) = \langle\mathbf{V}(t), d\mathbf{X}_3(t)\rangle_F = \langle\mathbf{V}(t), (D_{\mathbf{A}(t)}\exp)(d\mathbf{A}(t))\rangle_F = \langle(D_{\mathbf{A}(t)}\exp)(\mathbf{V}_s(t)), d\mathbf{A}(t)\rangle_F,$$

where again $\mathbf{V}_s(t) = \frac{1}{2}(\mathbf{V}(t) + \mathbf{V}(t)^\top)$. Then we obtain

$$\begin{aligned} df(t) &= \langle(D_{\mathbf{A}(t)}\exp)(\mathbf{V}_s(t)), \frac{\alpha}{H}(D_{\mathbf{X}_1^h(t)}\log)(d\mathbf{X}_1^h)\rangle_F \\ &= \frac{\alpha}{H}\langle(D_{\mathbf{X}_1^h(t)}\log)((D_{\mathbf{A}(t)}\exp)(\mathbf{V}_s(t))), d\mathbf{X}_1^h\rangle_F. \end{aligned}$$

Finally, we obtain $df(t) = \langle\frac{2\alpha}{H}(D_{\mathbf{X}_1^h(t)}\log)((D_{\mathbf{A}(t)}\exp)(\mathbf{V}_s(t)))\mathbf{W}_1^h\mathbf{X}(t), d\mathbf{W}_1^h\rangle_F$, and therefore the **Euclidean** gradient w.r.t. $\mathbf{W}_1^h$ is

$$\nabla_{\mathbf{W}_1^h}f(t) = \frac{2\alpha}{H}(D_{\mathbf{X}_1^h(t)}\log)((D_{\mathbf{A}(t)}\exp)(\mathbf{V}_s(t)))\mathbf{W}_1^h\mathbf{X}(t). \tag{18}$$

LC metric Under the LC metric, for a perturbation $\Delta\mathbf{W}$ in $\mathbf{W}_1^h$, we have $d\mathbf{X}_1^h(t) = \Delta\mathbf{W}\mathbf{X}(t)(\mathbf{W}_1^h)^\top + \mathbf{W}_1^h\mathbf{X}(t)\Delta\mathbf{W}^\top$, $d\mathbf{A}(t) = \frac{\alpha}{H}\sum_{h=1}^{H}(D_{\mathbf{X}_1^h(t)}\Phi)(d\mathbf{X}_1^h(t))$ and $d\mathbf{X}_3(t) = (D_{\mathbf{A}(t)}\Phi^{-1})(d\mathbf{A}(t))$. Let $\mathbf{V}(t)$ be an "upstream gradient", and write $f(t) = \langle\mathbf{V}(t), \mathbf{X}_3(t)\rangle_F$. Let $(D_{\mathbf{A}(t)}\Phi^{-1})^*$ be the adjoint operator of $(D_{\mathbf{A}(t)}\Phi^{-1})$, and $(D_{\mathbf{X}_1^h(t)}\Phi)^*$ the adjoint operator of $(D_{\mathbf{X}_1^h(t)}\Phi)$. Then we have

$$\begin{aligned} df(t) &= \frac{\alpha}{H}\langle(D_{\mathbf{X}_1^h(t)}\Phi)^*((D_{\mathbf{A}(t)}\Phi^{-1})^*(\mathbf{V}_s(t))), d\mathbf{X}_1^h\rangle_F \\ &= \langle\frac{2\alpha}{H}(D_{\mathbf{X}_1^h(t)}\Phi)^*((D_{\mathbf{A}(t)}\Phi^{-1})^*(\mathbf{V}_s(t)))\mathbf{W}_1^h\mathbf{X}(t), d\mathbf{W}_1^h\rangle_F, \end{aligned}$$

where $\mathbf{V}_s(t) = \frac{1}{2}(\mathbf{V}(t) + \mathbf{V}(t)^\top)$. Then the **Euclidean** gradient w.r.t. $\mathbf{W}_1^h$ is

$$\nabla_{\mathbf{W}_1^h}f(t) = \frac{2\alpha}{H}(D_{\mathbf{X}_1^h(t)}\Phi)^*((D_{\mathbf{A}(t)}\Phi^{-1})^*(\mathbf{V}_s(t)))\mathbf{W}_1^h\mathbf{X}(t). \tag{19}$$

We below give the explicit adjoints $(D_{\mathbf{A}(t)}\Phi^{-1})^*$ and $(D_{\mathbf{X}_1^h(t)}\Phi)^*$ for the log-Cholesky coordinates.

Let $\mathbf{L}(t) = \mathcal{L}(\mathbf{X}_3(t))$ denote the Cholesky factor of $\mathbf{X}_3(t)$. Define $\mathbf{B}(t) = \mathbf{L}(t)^\top\mathbf{V}_s(t)$; The adjoint $(D_{\mathbf{A}(t)}\Phi^{-1})^*(\mathbf{V}_s(t)) \in \mathbb{R}^{m_1(m_1+1)/2}$ is the vector of the lower-diagonal entries of the following matrix:

$$2\mathrm{diag}(\mathbf{L}(t))\mathrm{diag}(\mathbf{B}(t)) + \lfloor\mathbf{B}(t)\rfloor + \lfloor\mathbf{B}(t)^\top\rfloor.$$

Let $\mathbf{L}^h(t) = \mathcal{L}(\mathbf{X}_1^h(t))$ denote the Cholesky factor of $\mathbf{X}_1^h(t)$ for $h = 1, \ldots, H$. Given a coordinate covector $\boldsymbol{g}(t) \in \mathbb{R}^{m_1(m_1+1)/2}$, build a lower-triangular $\mathbf{H}^h(t)$ by

$$[\mathbf{H}^h(t)]_{i,i} = \frac{g_{\kappa(i,i)}(t)}{[\mathbf{L}^h(t)]_{i,i}}, \;\; [\mathbf{H}^h(t)]_{i,j} = g_{\kappa(i,j)}(t) \; (i > j), \;\; [\mathbf{H}^h(t)]_{i,j} = 0 \; (i < j),$$

where κ is the indexing function. Let Ψ be the lower-symmetrizer. Then

$$(D_{\mathbf{X}_1^h(t)}\Phi)^*(\boldsymbol{g}(t)) = \frac{1}{2}\mathbf{L}^h(t)^{-\top}(\Psi(\mathbf{L}^h(t)^\top \mathbf{H}^h(t)) + \Psi(\mathbf{L}^h(t)^\top \mathbf{H}^h(t))^\top)\mathbf{L}^h(t)^{-1}.$$

F CONGRUENCE MATRIX REGULARIZATION

Without additional regularization on the congruence weight matrices $\{\mathbf{W}_1^h\}_{h=1}^H$ and $\{\mathbf{W}_4^h\}_{h=1}^H$, different heads are likely to focus on the same dominant patterns in the functional data. To encourage the heads to learn complementary features, we introduce a row-space decorrelation penalty.

For the matrices $\{\mathbf{W}_1^h\}_{h=1}^H$, a natural choice is to penalize the overlap between their row spaces. Specifically, for each pair $1 \le h < k \le H$, we consider

$$\|\mathbf{W}_1^h(\mathbf{W}_1^k)^\top\|_F^2,$$

which quantifies the degree of alignment between the row spaces of $\mathbf{W}_1^h$ and $\mathbf{W}_1^k$. Summing over all pairs gives the total pairwise-overlap penalty:

$$\mathcal{L}_{\text{pair}} = \sum_{1 \le h < k \le H} \|\mathbf{W}_1^h(\mathbf{W}_1^k)^\top\|_F^2.$$

To reduce the computational cost, we exploit the fact that each matrix $\mathbf{W}_1^h$ is constrained to be full row rank with orthonormal rows, that is, $\mathbf{W}_1^h(\mathbf{W}_1^h)^\top = \mathbf{I}_{m_1}$. For each head h, we define

$$\mathbf{P}_h = (\mathbf{W}_1^h)^\top \mathbf{W}_1^h \in \mathbb{R}^{m \times m},$$

which is the orthogonal projector onto the row space of $\mathbf{W}_1^h$. Using $\|\mathbf{W}_1^h(\mathbf{W}_1^k)^\top\|_F^2 = \operatorname{tr}(\mathbf{P}_h\mathbf{P}_k)$, we can rewrite the total pairwise-overlap penalty as

$$\mathcal{L}_{\text{pair}} = \sum_{1 \le h < k \le H} \operatorname{tr}(\mathbf{P}_h\mathbf{P}_k) = \frac{1}{2}\left(\|\sum_{h=1}^H \mathbf{P}_h\|_F^2 - \sum_{h=1}^H \|\mathbf{P}_h\|_F^2\right).$$

Moreover, for each projector $\mathbf{P}_h$,

$$\|\mathbf{P}_h\|_F^2 = \operatorname{tr}(\mathbf{P}_h^2) = \operatorname{tr}(\mathbf{P}_h) = m_1.$$

Hence,

$$\mathcal{L}_{\text{pair}} = \frac{1}{2}\left(\|\sum_{h=1}^H \mathbf{P}_h\|_F^2 - Hm_1\right).$$

Therefore, minimizing $\mathcal{L}_{\text{pair}}$ is equivalent to minimizing $\|\sum_{h=1}^H \mathbf{P}_h\|_F^2$.

When $Hm_1 > m$, the row spaces of all heads cannot be mutually orthogonal. In this overcomplete regime, a more appropriate objective is to distribute the row spaces as uniformly as possible, which corresponds to encouraging the sum of the projection matrices to be close to an isotropic target:

$$\sum_{h=1}^H \mathbf{P}_h \approx \frac{Hm_1}{m}\mathbf{I}_m.$$

Accordingly, we define the following diversity regularizer:

$$\mathcal{L}_{\text{div}} = \lambda \left\| \sum_{h=1}^H (\mathbf{W}_1^h)^\top \mathbf{W}_1^h - \frac{Hm_1}{m}\mathbf{I}_m \right\|_F^2,$$

where $\lambda > 0$ is a tuning parameter. The same regularization is also applied to $\{\mathbf{W}_4^h\}_{h=1}^H$:

$$\mathcal{L}_{\text{div}} = \lambda \left\| \sum_{h=1}^H (\mathbf{W}_4^h)^\top \mathbf{W}_4^h - \frac{Hm_2}{m_1}\mathbf{I}_{m_1} \right\|_F^2.$$

G SUPPLEMENTARY EXPERIMENTAL DETAILS AND RESULTS

G.1 DETAILS ON THE CLUSTERING BENCHMARK METHODS AND THEIR CONFIGURATION

TSRVF In (Dai et al., 2020), trajectory dissimilarity is measured in the Transported Square-Root Vector Field (TSRVF) space, with temporal reparameterization used to factor out inter-subject differences in execution rate. The authors adopted a particular Riemannian metric on the SPD manifold, which provides closed-form expressions for the key geometric operations they need (e.g., parallel transport). When the matrix dimension m is large (e.g., $m > 100$), efficient comparison of covariance trajectories becomes computationally impractical. Therefore, Dai et al. (2020) apply dimension reduction before trajectory comparison. We adopt the same approach here and set the reduced dimension to 8. The other hyperparameter is the number of optimization iterations which is set to 30.

GeoAtt The network feeds each SPD matrix $\mathbf{X}(t)$ into a manifold-aware CNN to produce a lower-dimensional SPD matrix $\mathbf{Z}(t)$, and the sequence $\{\mathbf{Z}(t)\}$ is then processed by a manifold-aware GRU. We take the hidden state from the GRU as the latent embedding of the whole trajectory $\{\mathbf{X}(t)\}$. The hyperparameters and their values are: epochs $= 100$, latent dimension $= 8$, Adam learning rate $= 0.001$ (Euclidean parameters), Stiefel manifold learning rate $= 0.005$ (BiMap weights), and batch size $= 16$.

SPDNet SPDNet learns a latent vector embedding $\mathbf{x}(t)$ for each input SPD matrix $\mathbf{X}(t)$. Given a sequence of SPD matrices $\{\mathbf{X}(t_1), \mathbf{X}(t_2), \mathbf{X}(t_3), \ldots\}$, we average the corresponding latent embeddings $\{\mathbf{x}(t_1), \mathbf{x}(t_2), \mathbf{x}(t_3), \ldots\}$, and use this mean vector as the latent representation of the entire SPD matrix-valued function. The hyperparameter settings are as follows: training epochs $= 100$, latent dimension $= 8$, Adam learning rate $= 0.001$ (Euclidean parameters), Stiefel manifold learning rate $= 0.005$ (BiMap weights), and batch size $= 16$. The intermediate BiMap layer dimensions are determined automatically by $m_1 = \max(4, \lfloor m/2 \rfloor)$ and $m_2 = \max(3, \lfloor m_1/2 \rfloor)$.

PGA The authors in (Fletcher et al., 2004) introduced principal geodesic analysis (PGA), which is essentially a nonlinear generalization of PCA to manifold-valued data. In practice, PGA is implemented by mapping the manifold-valued data to the tangent space at the intrinsic mean via the Logarithm map and then performing standard PCA on the resulting tangent vectors. In our approach, we perform standard functional PCA in the tangent space, and then apply k-means on the principal component score vectors. We adopt the log-Euclidean metric and retain 10 principal components.

PDM We compute the pairwise distances between functions using the affine-invariant metric and then apply k-medoids on the resulting distance matrix. The number of random initializations for the clustering method is 10.

DC DC is the standard MLP autoencoder for deep clustering (DC). To adapt it to functional data, we make the input layer contain $m(m+1)/2$ nodes, where each node receives a T-dimensional input vector (corresponding to one $X_{ij}(t)$ observed over T time points). The hyperparameter settings are as follows: training epochs $= 100$, latent dimension $= 8$, Adam learning rate $= 0.001$, and batch size $= 16$. The encoder consists of 1D convolutional layers with channel sizes $\mathrm{tri} \to 64 \to 32$, followed by a fully connected layer that produces the latent representation.

MatFAE The algorithm repository documentation provides a clear explanation of all hyperparameters and their default values.

For all four deep learning models, including GeoAtt, SPDNet, DC, and MatFAE, we used Bayesian optimization implemented in Optuna for hyperparameter tuning. In all the experiments below, MatFAE training and competitive methods computations were performed on a 2.5 GHz Intel Xeon Gold 6548Y+ CPU (32 cores) and an NVIDIA L40S GPU with 46 GB of VRAM.

G.2 DETAILS ON THE CLASSIFICATION BENCHMARK METHODS AND THEIR CONFIGURATION

Shared protocol for all SPD-native methods (SPD-SRU, BW-norm, MatFAE-C): internal 80/20 split with early stopping on validation balanced accuracy (patience 25), class-weighted cross-entropy loss, Adam optimizer on Euclidean weights combined with a manual Stiefel-manifold step (learning

rate 5×10^{-3}) on BiMap/orthogonal weights, gradient clipping at 5.0, and an SPD floor of $\varepsilon = 10^{-4}$. STAGIN, which does not operate on the SPD manifold, uses the same early-stopping and loss protocol but standard Euclidean Adam throughout (no Stiefel step or SPD floor). All architecture designs and default hyperparameters (learning rate, weight decay, batch size, dropout, epoch budget) otherwise follow each method's original publication, as detailed below.

SPD-SRU Each frame is reduced to an 8×8 SPD matrix via BiMap, then passed through one SPD-SRU recurrent cell with five time scales $\alpha = [0.01, 0.25, 0.5, 0.9, 0.99]$ (multi-scale log-Euclidean-weighted running SPD means, orthogonal recurrence, learnable scale/output/blend gates), followed by LogEig→vech→head. Batch size 40, weight decay 0, ≤200 epochs (early-stopped), dropout 0.1.

STAGIN Per-frame correlation graphs are thresholded to a top-30% binary adjacency and passed through a 4-layer GIN (sum aggregation, hidden dimension 128) with per-layer SERO readouts summed, sinusoidal temporal positional encoding, one-head temporal self-attention, and a summed CLS token over time; a linear head produces the final prediction. Learning rate 5×10^{-4}, weight decay 10^{-5}, batch size 3, 100 epochs, dropout 0.5.

BW-norm Standard SPDNet-style BiMap layer → BW batch normalization (running BW barycenter, momentum 0.1, $\varepsilon = 10^{-5}$, geodesic running mean; centering in the BW log-chart, scaling by BW variance, with a learnable SPD bias) → ReEig → LogEig → linear head; per-frame outputs are averaged before the final head. Learning rate 2.5×10^{-3}, weight decay 5×10^{-2}, batch size 30, ≤200 epochs, dropout 0.1.

MatFAE-C We retain the `MatFAE` encoder backbone, remove the decoder, and append a linear classification head to the latent representation $\boldsymbol{x} = \mathcal{E}(\mathcal{F}(\mathbf{X}))$. The objective combines class-weighted cross-entropy with the same regularization terms used in the unsupervised model, namely congruence matrix regularization, the orthogonality penalty on functional weights, and the roughness penalty on functional weights; class weights are computed within each training fold using `sklearn.utils.class_weight.compute_class_weight`. Evaluation uses a class-stratified outer five-fold split. Euclidean parameters use Adam with learning rate 10^{-3} and weight decay 10^{-5}, batch size 16, latent dimension 16, and dropout 0.1 before the classification head. Unless otherwise stated, all backbone hyperparameters $(H, m_1, m_2, p_1, p_2, \alpha, \alpha_{\min}, \text{basis } K)$, encoder and MLP widths, and loss weights are set to their default values.

G.3 Details on the simulation scenarios and additional experimental results

Each synthetic dataset consists of 100 functional trajectories, each defined on the manifold $\mathcal{S}_+^{48}$. The matrix dimension is determined by the 48 cortical regions in the Harvard-Oxford atlas. Below, let q denote the number of time points at which the function $\mathbf{X}(t)$ is evaluated over the interval $[t_0, t_1]$, and k the true number of clusters.

The benchmark is organized as an ablation ladder, where each synthetic dataset isolates a single type of structural signal. Consequently, failure on a given rung suggests that the method does not possess the feature-extraction mechanism required to capture that specific structure.

A $(q = 40, k = 2)$: The data generation process is as follows. First, we define two well-separated SPD states, A and B. For Group 0, each trajectory starts in state A and switches to state B halfway through, producing an A→B pattern. For Group 1, each trajectory starts in state B and switches to state A halfway through, producing a B→A pattern. To introduce between-subject variability, we add random jitter of up to ±3 windows to the transition point, together with tangent-space noise. *Rung: temporal ordering.*

B $(q = 40, k = 2)$: We define three SPD states, A, B, and C. Every trajectory begins and ends in state A. The two groups differ in the intermediate state they visit: Group 0 passes through state B, while Group 1 passes through state C. *Rung: static identity.*

C $(q = 120, k = 3)$: We follow the HMM-based dFC construction of (Vidaurre et al., 2017). In particular, we consider three brain states and assign each group a different Markov transition matrix, such that the corresponding stationary dwell-time ratios are 75/15/10, 50/30/20, and 25/35/40, respectively. *Rung: dwell time.*

D $(q = 60, k = 2)$: Both groups switch between the same two states, A and B, following piecewise-constant trajectories. Group 0 has a switching period of approximately $q/4$, while Group 1 has a shorter switching period of approximately $q/12$. As a result, both groups share the same visited-state set and stationary proportions, but differ in how frequently they transition between states. *Rung: transition frequency.*

E $(q = 30, k = 2)$: Each trajectory is generated by sampling a symmetric-matrix-valued Gaussian process in the tangent space with squared-exponential kernel

$$\kappa_\ell(t, t') = \exp\left(-\frac{(t - t')^2}{2\ell^2}\right),$$

and then mapping it to the SPD manifold through $\mathbf{X}(t) = \exp(\mathbf{Y}(t))$. Group 0 uses $\ell = 12$, while Group 1 uses $\ell = 2$. The cluster means are asymptotically the same, so the groups are distinguished only by trajectory smoothness. *Rung: smoothness.*

F $(q = 20, k = 3)$ For each time point, we sample $\mathbf{X}(t) \sim \mathcal{W}_m(\mathbf{V}, v_c)/v_c$, with common scale matrix $\mathbf{V} = \mathbf{I}$ and group-specific degrees of freedom $v_c \in \{150, 75, 50\}$. Because the mean is the same for all groups, the groups differ only by their concentration, as reflected in quantities such as the log-determinant and trace. *Rung: Wishart concentration.*

G $(q = 30, k = 2)$: We generate group-specific neural source signals and convolve them with the canonical double-gamma hemodynamic response function (HRF) (Glover, 1999). We then construct the functional connectivity trajectory $\mathbf{X}(t)$ by applying sliding-window Pearson correlation to the simulated signals (Allen et al., 2014). *Rung: real HRF-blurred functional connectivity.*

H $(q = 30, k = 2)$: Following (Arsigny et al., 2007; Pennec et al., 2006), we consider two groups that share the same endpoints, $\mathbf{X}_0$ and $\mathbf{X}_1$. The geodesic connecting these two points on the SPD manifold is given by

$$\gamma(t) = \mathbf{X}_0^{1/2} \exp(t \log(\mathbf{X}_0^{-1/2} \mathbf{X}_1 \mathbf{X}_0^{-1/2})) \mathbf{X}_0^{1/2}.$$

For Group 0, the trajectory follows the geodesic under the identity warp: $\tau \mapsto \tau$. For Group 1, the trajectory follows the same geodesic under a triangle-bounce warp: $\tau \mapsto 2\tau$ for $\tau \leq 1/2$, and $\tau \mapsto 2(1 - \tau)$ for $\tau > 1/2$. Since the two warps have the same mean $(1/2)$ and variance $(1/12)$, the Jensen gap is zero. Thus, the groups are distinguished only by temporal progression along the geodesic. *Rung: trajectory direction.*

I $(q = 60, k = 2)$: We generate the data using a two-level hierarchical HMM (Fine et al., 1998). The outer state, taking values in $\{O_1, O_2\}$, evolves slowly over time, while the inner state, taking values in $\{I_1, I_2, I_3\}$, evolves conditional on the current outer state. The two groups differ in the coupling structure between the outer and inner states. *Rung: multi-scale dynamics.*

J $(q = 50, k = 2)$**:** The two groups are anchored to meta-analytic brain maps from the Neurosynth database (Yarkoni et al., 2011), accessed through `NiMARE` (Salo et al., 2023). We selected six cognitive terms and divided them into two concept families: an attentional-control family (*attention*, *executive*, *inhibition*) and a memory family (*memory*, *recall*, *encoding*). For each term t, we identified all Neurosynth studies whose abstract-level TF-IDF loading for that term exceeded 10^{-3}. We then computed a multilevel kernel density analysis map by placing a 10 mm spherical kernel at each reported peak coordinate and summing the resulting kernels across studies. Each term-specific density map was parcellated into the 48 Harvard-Oxford cortical regions, producing a regional score vector $\mathbf{s}(t) \in \mathbb{R}^{48}$. This vector was mean-centred and unit-normalised to obtain $\mathbf{u}(t)$, and then converted into an SPD anchor matrix using $\mathbf{X}(t) = \epsilon\mathbf{I} + \mathbf{u}(t)\mathbf{u}(t)^\top$, where ϵ was chosen so that $\mathbf{X}(t)$ had condition number 7.5. For each synthetic trajectory, the sequence was divided into three approximately equal temporal segments. Each group cycled through the three Neurosynth-derived anchors from its own concept family, with a random ± 3-window jitter applied to each segment boundary. Additional log-Euclidean tangent-space perturbations were added at both the subject and window levels.

Table 3 shows a pattern broadly consistent with the AMI results. MatFAE achieves the highest mean ARI in seven scenarios (A, C, D, E, F, H and J), with especially large advantages in the more

Table 3: ARI scores for the ten simulation scenarios. The table reports the mean (top row) and standard deviation (bottom row) of the scores over 100 repetitions.

Scenario (*rung*)	TSRVF	GeoAtt	SPDNet	PGA	DC	PDM	MatFAE
A (*temporal ordering*)	0.999	-0.001	-0.000	-0.001	0.990	-0.000	**1.000**
	0.009	0.011	0.013	0.012	0.101	0.012	0.000
B (*static identity*)	0.010	0.028	0.043	**0.999**	0.007	0.961	**0.980**
	0.029	0.059	0.070	0.007	0.028	0.119	0.009
C (*dwell time*)	0.295	0.437	0.429	0.439	0.434	0.440	**0.541**
	0.069	0.082	0.065	0.069	0.083	0.069	0.032
D (*transition frequency*)	0.279	-0.002	0.000	-0.000	0.272	0.001	**0.286**
	0.062	0.011	0.014	0.013	0.107	0.011	0.052
E (*smoothness*)	0.064	0.061	0.011	0.063	0.012	0.000	**0.808**
	0.062	0.043	0.027	0.052	0.016	0.000	0.078
F (*Wishart concentration*)	0.077	0.999	0.439	0.672	0.128	0.535	**1.000**
	0.045	0.005	0.080	0.072	0.062	0.075	0.000
G (*real HRF*)	-0.001	0.000	0.004	**1.000**	0.981	0.170	0.960
	0.012	0.015	0.020	0.000	0.115	0.124	0.018
H (*trajectory direction*)	**1.000**	0.192	0.038	0.012	**1.000**	0.001	**1.000**
	0.000	0.092	0.042	0.028	0.000	0.013	0.000
I (*multi-scale dynamics*)	**0.097**	0.034	0.035	0.033	0.046	0.034	0.085
	0.072	0.046	0.043	0.043	0.053	0.044	0.064
J (*Neurosynth*)	0.000	0.282	0.791	**1.000**	0.784	**1.000**	**1.000**
	0.013	0.267	0.121	0.000	0.343	0.000	0.000

challenging settings C and E. The gain in the smoothness scenario E is particularly striking: MatFAE reaches 0.808, whereas all competing methods remain near zero. It also performs best in F and ties for perfect recovery in H and J.

The baseline methods again show more specialized behavior. TSRVF and DC perform very well in the temporally driven scenarios A and H, while PGA is strongest in the static identity and real-HRF scenarios B, J and G. Scenario I remains the most difficult overall, with all methods achieving low ARI values; unlike the AMI results, TSRVF attains the highest mean score there, although the margin over MatFAE is small. Overall, the ARI results reinforce the conclusion that MatFAE is the most robust method across the benchmark, with its clearest advantage appearing in the harder scenarios.

G.4 Details on the real datasets and additional experimental results

All six resting-state fMRI cohorts were obtained from publicly available preprocessed releases. We retained the preprocessing provided by each consortium, including motion correction, T1 coregistration, spatial normalization to MNI space, and, where applicable, upstream confound regression. To improve comparability across datasets, we then applied a common post-hoc processing pipeline where the required data were available.

Unless stated otherwise, the post-hoc pipeline included nuisance regression using the full Friston-24 motion model, 5-6 aCompCor components, and volume-level scrubbing for frames with framewise displacement greater than $0.5\,\mathrm{mm}$. Scrubbing was implemented either as one-hot spike regressors or through sample-mask exclusion, depending on the confound information provided with each release. We then applied temporal band-pass filtering at 0.01-$0.1\,\mathrm{Hz}$, linear and quadratic detrending, and per-ROI z-scoring.

When native NIfTI images were available, we parcellated the data using the 100-parcel, 7-network Schaefer atlas in MNI $2\,\mathrm{mm}$ space. When only pre-extracted ROI time series were provided, we used the atlas distributed with the corresponding release. Time-varying functional connectivity was estimated using Ledoit-Wolf shrinkage covariance matrices computed within $30\,\mathrm{s}$ sliding windows with a $4\,\mathrm{s}$ step. This produced one SPD-valued covariance trajectory for each subject.

Dataset-specific deviations from this common pipeline are described below. Here, n denotes the number of retained subjects, q the number of sliding windows, m the final ROI dimension after applying a common non-zero-variance mask, and k the number of diagnostic classes.

CNP ($n = 257,\ q = 57,\ m = 100,\ k = 4$) **(Bilder et al., 2020)** : The Consortium for Neuropsychiatric Phenomics (CNP) dataset was obtained from OpenNeuro under accession number `ds000030`. The cohort comprises healthy controls ($n = 120$), individuals with schizophrenia ($n = 48$), bipolar disorder ($n = 49$), and ADHD ($n = 40$). Each participant has one resting-state fMRI run with TR $= 2\,\mathrm{s}$ and 152 volumes. We used the fMRIPrep-v0.4.4 preprocessed release (Gorgolewski et al., 2017), which provides motion-corrected BOLD images spatially normalized to the MNI152NLin2009cAsym template (the asymmetric 2009c nonlinear MNI152 standard space) at 2 mm resolution. The release also includes a 24-column confound TSV file containing 6 motion parameters, 6 aCompCor components, 6 tCompCor components, white-matter signal, global signal, framewise displacement, and DVARS.

We applied a standard preprocessing pipeline consisting of Friston-24 motion regression, 6 aCompCor components, FD $> 0.5\,\mathrm{mm}$ spike regression, and temporal band-pass filtering. The data were then parcellated in MNI space using the 100-parcel Schaefer atlas. Ledoit-Wolf covariance matrices were computed over 30 s sliding windows, corresponding to 15 TRs, with a step size of 2 TRs. Four subjects were excluded because more than 50% of their volumes exceeded the FD threshold.

COBRE ($n = 144,\ q = 68,\ m = 100,\ k = 2$) : The COBRE (RRID:SCR_010482) dataset was obtained from the International Neuroimaging Data-sharing Initiative. The cohort includes individuals with schizophrenia ($n = 70$) and healthy controls ($n = 74$), giving a total sample size of 144 participants. Each resting-state scan was acquired with TR $= 2\,\mathrm{s}$ and contains 150 volumes. The data were preprocessed upstream by Bellec and colleagues using the NIAK pipeline (v0.17). This preprocessing produced BOLD images in MNI152NLin2009a space at 6 mm isotropic resolution, together with per-subject confound TSV files. These confound files include 6 motion parameters, framewise displacement, a scrub flag, 6 slow-drift discrete cosine transform basis functions, white-matter and ventricle mean signals, and 5 aCompCor components.

Importantly, these confound variables are provided with the data but were not regressed out in the released BOLD time series. We therefore applied our common nuisance-regression and filtering pipeline, including Friston-24 motion regressors, 5 aCompCor components, white-matter and ventricle signals, one-hot spike regressors for volumes marked with scrub = 1, and temporal band-pass filtering. The resulting time series were parcellated in MNI space using the 100-parcel Schaefer atlas. Time-varying functional connectivity was then estimated by computing Ledoit-Wolf covariance matrices within 30 s sliding windows, corresponding to 15 TRs, with a step size of 2 TRs.

ADHD-200 ($n = 138,\ q = 57,\ m = 100,\ k = 2$) (The ADHD-200 Consortium, 2012; Bellec et al., 2017): The ADHD-200 dataset is a multi-site cohort released by the ADHD-200 Consortium and preprocessed by the Preprocessed Connectomes Project. We used the CPAC-preprocessed version distributed at `s3://fcp-indi/data/Projects/ADHD200/Outputs/cpac/`, with the configuration `pc10.linear1.wm0.global0.motion1.quadratic1.gm0.compcor1.csf0`. The cohort includes typically developing controls ($n = 88$) and individuals with ADHD ($n = 50$). Data were acquired across 10 sites, with site-specific TR values of 1.5, 1.96, 2.0, or 2.5 s. In this preprocessing stream, 6 motion parameters, 5 aCompCor components, and linear and quadratic trends were already regressed from the BOLD time series. ROI time series were then extracted using the 100-parcel Schaefer atlas.

Because only the extracted ROI time series are available, rather than the raw BOLD images or confound TSV files, our post-hoc preprocessing was limited to temporal band-pass filtering using the site-specific TR. We then estimated time-varying functional connectivity by computing Ledoit-Wolf covariance matrices within sliding windows. Additional Friston-24 motion regression or volume scrubbing could not be applied retrospectively. Subjects were excluded if their scan duration satisfied $q \cdot \mathrm{TR} < 240\,\mathrm{s}$, or if more than 5% of parcels had zero variance. This led to the exclusion of one outlier, `sub-0015038`, and allowed the full 100-ROI Schaefer atlas to be retained across the cohort.

ABIDE-I ($n = 846,\ q = 57,\ m = 101,\ k = 2$) : The Autism Brain Imaging Data Exchange I (ABIDE I) dataset (Di Martino et al., 2014) is a multi-site resting-state fMRI cohort comprising healthy controls ($n = 455$) and individuals with autism spectrum disorder (ASD; $n = 391$). Data were acquired across 17 sites, with TR values ranging from 1.5 to 3.0 s. We used the CPAC-preprocessed `nofilt_noglobal` release from the Preprocessed Connectomes Project, distributed at `s3://fcp-indi/data/Projects/ABID_Initiative/Outputs/cpac/nofilt_noglobal/`. In this preprocessing stream, 24 Friston motion parameters, 5 aCompCor components, and linear and quadratic trends were regressed from the BOLD data. As indicated by the `nofilt` label, temporal band-pass filtering was not applied upstream; volume scrubbing was also not performed.

Because Schaefer-atlas time series are not provided in the PCP release for ABIDE I, we used the shipped Harvard-Oxford ROI time series directly. Our post-hoc preprocessing therefore consisted of temporal band-pass filtering using the site-specific TR, followed by estimation of time-varying functional connectivity using Ledoit-Wolf covariance matrices within sliding windows. After applying a common non-zero-variance mask across all subjects, 101 of the 111 Harvard-Oxford parcels were retained.

TCP ($n = 241,\ q = 80,\ m = 100,\ k = 2$) **(Chopra et al., 2025)** : The Transdiagnostic Connectome Project (TCP) dataset was obtained from OpenNeuro under accession number `ds005237`. The cohort includes healthy general-population controls ($n = 92$) and transdiagnostic patients ($n = 149$), aged 18-70 years, recruited at two acquisition sites: Yale University and McLean Hospital. Each participant has four resting-state fMRI runs, acquired with anterior–posterior and posterior–anterior phase-encoding directions (`task-restAP_run-01/02` and `task-restPA_run-01/02`). Each run was acquired with TR $= 0.8$ s, multiband factor 8, and 488 volumes. The OpenNeuro release provides time series after the consortium's HCP-style minimal preprocessing pipeline, including motion and distortion correction, T1 coregistration, spatial normalization to MNI space, CIFTI grayordinate projection, ICA-FIX automated noise classification and regression (Salimi-Khorshidi et al., 2014), global signal regression, and $1/2000$ Hz high-pass detrending. The released time series are parcellated using the 488-region CAB-NP atlas, comprising 360 Glasser cortical parcels (Glasser et al., 2016) and 128 subcortical CIFTI grayordinate parcels.

We retained the first anterior-posterior resting-state run (`task-restAP_run-01`) for each subject. Because the released data were already denoised and provided as parcellated time series, our post-hoc preprocessing was limited to temporal band-pass filtering at 0.01-0.1 Hz, followed by estimation of time-varying functional connectivity using Ledoit-Wolf covariance matrices. Covariance matrices were computed within 30 s sliding windows, corresponding to 38 TRs, with a step size of 5 TRs. To improve comparability with CNP, COBRE, and ADHD-200, we mapped the 360 Glasser cortical parcels to the 100-parcel Schaefer atlas using a row-stochastic voxel-overlap mapping computed in MNI152NLin2009cAsym 2 mm space; the 128 subcortical parcels were not used. Subjects were excluded if more than 50% of volumes exceeded the framewise-displacement threshold of 0.5 mm, based on the released `motion_FD/` files.

CAT-D ($n = 118,\ q = 49,\ m = 100,\ k = 2$) **(Sadeghi et al., 2022; Camp et al., 2024)** : The NIMH Characterization and Treatment of Adolescent Depression (CAT-D) dataset was obtained from OpenNeuro under accession number `ds004627`. The cohort includes healthy volunteers ($n = 50$) and adolescents with major depressive disorder ($n = 68$), all scanned at a single site on a GE Discovery MR750 3T scanner. Each participant has a multi-echo resting-state fMRI run with TR $= 2.5$ s and three echoes, with TE approximately 14.4, 28.8, and 43.2 ms. We retained only the second echo for parity with the other real datasets and restricted the analysis to the baseline session (`ses-v1`), so that each subject contributed a single resting-state trajectory.

Unlike the other datasets, CAT-D does not have a publicly available upstream-preprocessed release. We therefore preprocessed the data using fMRIPrep v23.2.0 (Esteban et al., 2019), which produced motion-corrected BOLD images spatially normalized to the MNI152NLin2009cAsym template at 2 mm resolution, together with per-subject confound TSV files. We then applied our post-hoc preprocessing pipeline, including Friston-24 motion regression, 6 anatomical aCompCor components, white-matter and CSF mean signals,

volume censoring based on FD $> 0.5\,\mathrm{mm}$ and standardized DVARS > 1.5, and temporal band-pass filtering at 0.01-$0.1\,\mathrm{Hz}$. The resulting time series were parcellated in MNI space using the 100-parcel Schaefer atlas. Time-varying functional connectivity was estimated by computing Ledoit-Wolf covariance matrices within 30 s sliding windows, corresponding to 12 TRs, with a step size of 2 TRs.

Table 4: The ARI scores for the six real datasets.

Dataset	TSRVF	GeoAtt	SPDNet	PGA	DC	PDM	MatFAE
CNP	0.057	0.017	0.015	0.039	0.018	0.001	**0.128**
COBRE	0.087	0.071	0.071	0.056	0.018	0.056	**0.169**
ADHD-200	0.037	0.005	0.006	0.040	0.075	0.078	**0.109**
ABIDE-I	0.014	0.001	0.001	0.000	0.005	0.001	**0.038**
TCP	0.036	0.002	0.013	0.028	0.005	0.025	**0.053**
CAT-D	0.026	0.031	0.007	0.056	0.000	0.039	**0.085**

`MatFAE` achieves the highest ARI on all six cohorts, indicating a consistent advantage over the competing methods. The absolute ARI values, however, remain modest, which reflects the difficulty of unsupervised diagnostic discovery from resting-state fMRI. Several patterns are worth noting. First, the deep-learning baselines, especially GeoAtt and SPDNet, perform close to chance on several cohorts, including CNP, ABIDE-I, TCP, and CAT-D. This suggests that parameter-heavy SPD encoders may be vulnerable to weak disease signal, limited sample size, site effects, and preprocessing variability. Second, the strongest competing baselines differ across cohorts: TSRVF is the strongest baseline on CNP and COBRE, PDM and DC are strongest on ADHD-200, and PGA is strongest on CAT-D. Nevertheless, `MatFAE` remains consistently ahead of these cohort-specific competitors, with the largest margins observed on COBRE and CNP. This suggests that the combination of geometric layers and functional feature extraction provides a more stable representation than relying only on temporal alignment, tangent-space summaries, or static geometric structure. Read together with the simulation study, the real-data results support the same conclusion. Individual baselines are competitive in specific regimes, such as TSRVF and DC for temporal ordering, PGA for static identity or HRF-blurred connectivity, and GeoAtt for Wishart concentration. By contrast, `MatFAE` performs strongly across multiple regimes and is particularly effective on dwell time, transition frequency, and smoothness, which are relevant to resting-state dynamic connectivity. The real fMRI results therefore show that `MatFAE` is more discriminative than the tested alternatives, while the low absolute ARI values also highlight that unsupervised recovery of DSM diagnostic labels from resting-state dynamic connectivity remains a challenging open problem.

Table 5: Balanced classification accuracy (mean $\pm$ std over five-fold CV) for MatFAE-C and three classification baselines. Chance-level accuracy is shown in parentheses next to each dataset name.

Dataset	SPD-SRU	STAGIN	BW-norm	MatFAE-C
CNP (0.25)	0.245 ± 0.032	0.287 ± 0.028	0.338 ± 0.032	0.453 ± 0.048
COBRE (0.5)	0.541 ± 0.036	0.638 ± 0.036	0.604 ± 0.033	0.741 ± 0.047
ADHD-200 (0.5)	0.519 ± 0.069	0.528 ± 0.042	0.571 ± 0.041	0.630 ± 0.072
ABIDE-I (0.5)	0.495 ± 0.024	0.587 ± 0.030	0.525 ± 0.034	0.628 ± 0.063
TCP (0.5)	0.505 ± 0.030	0.498 ± 0.034	0.498 ± 0.029	0.609 ± 0.067
CAT-D (0.5)	0.521 ± 0.040	0.542 ± 0.033	0.493 ± 0.050	0.642 ± 0.056

Classification results are reported in Table 5. The folds are not site-stratified and no ComBat harmonisation is applied; therefore, the reported results should be interpreted as within-distribution classification performance rather than evidence of out-of-site generalisation. The supervised variant `MatFAE-C` achieves the highest balanced classification accuracy on all six cohorts. Although the CNP accuracy appears lower in absolute terms, CNP is a four-class problem, for which chance-level balanced accuracy is 0.25 rather than 0.50; the CNP result therefore remains substantially above chance, and in fact represents the largest relative margin over the best baseline of any cohort (0.453 vs. 0.338 for BW-norm, a 34% relative improvement). Overall, the performance of `MatFAE-C` is comparable to conservative resting-state fMRI classification results reported on multisite psychiatric

datasets such as ABIDE-I and ADHD-200, where accuracies around 0.60-0.70 are commonly observed under realistic validation settings. However, it is below some highly optimized supervised studies on single binary datasets, such as COBRE schizophrenia classification, where substantially higher accuracies have been reported using task-specific feature engineering, feature selection, multimodal inputs, or deep supervised architectures. These comparisons should therefore be interpreted cautiously, as published results differ substantially in diagnostic task, number of classes, preprocessing, site handling, validation protocol, and whether external generalisation is evaluated.

G.5 Statistical significance tests and ablation study

We test the statistical significance for the reported improvements on both simulated and real data. For each synthetic rung, we test whether MatFAE (clustering) and MatFAE-C (classification) perform significantly above chance. Specifically, we apply a one-sided one-sample t-test to the scores across random seeds, using a chance level of 0 for adjusted ARI and 0.5 for balanced accuracy on the binary-class rungs (0.33 for the three-class rung). We report both the p-value and effect size (Cohen's d). The clustering analysis uses 100 seeds per rung, while the classification analysis uses 20 seeds (5×3 stratified CV). For the six real datasets, we apply the same procedure: For clustering, we set the number of clusters to the known number and test whether the mean ARI across 20 training seeds exceeds the chance level of 0. For classification, we test whether the mean balanced accuracy across the 15 outer test folds of the 5×3 stratified cross-validation exceeds the chance level of $1/K$, where K is the number of classes. This corresponds to 0.25 for the four-class CNP dataset and 0.5 for the remaining binary datasets. The test results are given in Table 6.

Table 6: Statistical significance of performance on the synthetic rungs and real datasets.

Scenario (*rung*)	p (MatFAE)	d	p (MatFAE-C)	d
A (*temporal ordering*)	<0.001	ceiling	<0.001	63.8
B (*static identity*)	<0.001	62.8	<0.001	8.3
C (*dwell time*)	<0.001	16.9	<0.001	4.5
D (*transition frequency*)	<0.001	5.1	<0.001	4.9
E (*smoothness*)	<0.001	10.4	<0.001	6.6
F (*Wishart concentration*)	<0.001	ceiling	<0.001	20.6
G (*real HRF*)	<0.001	10.8	<0.001	7.3
H (*trajectory direction*)	<0.001	ceiling	<0.001	50.9
I (*multi-scale dynamics*)	0.003	1.1	<0.001	19.3
J (*Neurosynth*)	<0.001	ceiling	<0.001	14.9
CNP	<0.001	12.7	<0.001	4.2
COBRE	<0.001	11.9	<0.001	5.1
ADHD-200	<0.001	5.8	0.008	1.8
ABIDE-I	<0.001	17.3	0.005	2.0
TCP	<0.001	3.3	0.011	1.6
CAT-D	<0.001	12.0	<0.001	2.5

We conduct a component-wise ablation, removing one module at a time (Table 7). Each model variant is evaluated on three representative simulation rungs (E, G, and J) and the mean ARI over 100 independent repetitions is reported. Across the three rungs, the most pronounced and recurring performance degradation occurs when the multi-head BiMap is reduced to a single head, the matrix-to-vector module $\mathcal{F}$ is replaced, or the functional layers are removed. These results indicate that all three components are critical to the performance of MatFAE.

We conduct a one-factor-at-a-time sensitivity analysis on the same three simulation rungs: E, G, and J. Each hyperparameter is varied while all others are held fixed. Table 8 reports the mean ARI over 100 independent repetitions.

The results reveal the following trends:

Number of heads H: Performance generally improves as additional heads capture distinct localized temporal patterns, before plateauing once the relevant temporal regimes are adequately represented.

Table 7: Component-wise ablation results for MatFAE on three simulation rungs. Each row indicates a component removed or simplified relative to the full model. Values are mean ARIs over 100 independent repetitions.

Module removed	E *smoothness*	G *real HRF*	J *Neurosynth*
– multi-head ($H \to 1$)	0.031	0.007	0.136
– functional layers	0.021	0.818	0.680
– module $\mathcal{F}$ (direct half-vectorization)	0.016	0.030	0.120
– module $\mathcal{F}$ (Riemannian logarithm vectorization)	0.025	-0.001	0.770
– second BiMap layer	0.674	0.866	0.939
– geodesic shrinkage ($\alpha \to 1$)	0.624	0.804	0.960
Full MatFAE	0.808	0.960	1.000

Table 8: One-factor-at-a-time hyperparameter sensitivity analysis on three simulation rungs. The number of heads (H), latent dimension (s), basis size (K), intermediate SPD dimension (m_1), and geodesic shrinkage parameter (α) are varied individually while the remaining hyperparameters are held fixed. Values are mean ARIs over 100 independent repetitions.

Hyper-parameter	E *smoothness*	G *real HRF*	J *Neurosynth*
H {1,2,4,8,16}	{.03, .11, .14, **.81**, .79}	{.01,.33,.51,.96,**.98**}	{.14, .28, .77, **1.00**, 1.00}
s {2,4,8,16,32}	{.64, **.79**, .78, .76, .78}	{.81,.89,**.95**,.93,.90}	{.96, .99, **1.00**, 1.00, .96}
K {6,10,15,20,25}	{.29, .40, **.82**, .79, .79}	{.64,.87,.93,.95,**.96**}	{.84, .94, **1.00**, 1.00, 1.00}
m_1 {16,24,32,48}	{.68, .72, **.81**, .80}	{.85, .88, **.96**, .96}	{.96, **1.00**, 1.00, 1.00}
α {.1,.25,.5,.75,.9}	{.54, .60, **.80**, .79, .62}	{.78,.84,**.94**,.86,.84}	{.94, .96, **1.00**, .99, 1.00}

Latent dimension s: Performance is relatively stable over a moderate range of latent dimensions. Small values can restrict representational capacity, whereas larger values provide no consistent benefit and may introduce redundant features.

Basis size K: Performance improves until the basis is sufficiently expressive, after which additional basis functions yield little or no benefit. Thus, K acts primarily as a minimum capacity requirement rather than following a larger-is-better relationship.

Intermediate SPD dimension m_1: Moderate values generally provide sufficient representational capacity. Increasing m_1 beyond this range produces only marginal improvements, indicating limited sensitivity once an adequate dimension is reached.

Geodesic shrinkage α: Intermediate values generally perform best. Because the pooling operation already introduces nonlinearity, geodesic shrinkage primarily acts as a regularizer; excessive or insufficient shrinkage can reduce performance.

G.6 Interpretability of MatFAE

Compared with existing complex deep learning models used in neuroscience, MatFAE combines a lightweight architecture with intrinsic interpretability. In particular, the temporal profile of the functional weights $\boldsymbol{W}^{(1)}(t)$ in the encoder functional layer indicates which slices of the SPD trajectory contribute most strongly to the latent representations. We illustrate this interpretability using the COBRE dataset by comparing statistics of $\boldsymbol{y}(t)$ with the temporal profile of $\boldsymbol{W}^{(1)}(t)$ learned by MatFAE under the configuration reported in Appendix G.4.

Profile and peak window Let $g(t) = \|\boldsymbol{W}^{(1)}(t)\|_F$ denote the Frobenius norm of the functional weight at time index t, and let $e(t) = \sqrt{\sum_{k=1}^{K} B_k(t)^2}$ denote the pointwise ℓ_2 envelope of the B-spline basis. Because clamped open-uniform B-splines satisfy $e(0) = e(T-1) = 1$, while the interior values are below unity, the raw norm $g(t)$ contains a deterministic boundary effect induced by the basis itself. To isolate the data-driven temporal preference of the encoder, we therefore use

the basis-corrected profile $\tilde{g}(t) = g(t)/e(t)$. The peak window $[t_1, t_2]$ is selected as the contiguous interior interval of width $\lfloor 0.20\,T \rceil$ centred at $\arg\max_t \tilde{g}(t)$, with two end-windows excluded on each side.

Cohort divergence in $\boldsymbol{y}(t)$ For each subject i, we compute $\boldsymbol{y}_i(t) \in \mathbb{R}^p$ and define two pointwise cohort-divergence summaries: the ℓ_2 norm of the cohort-mean difference, $\|\bar{\boldsymbol{y}}_P(t) - \bar{\boldsymbol{y}}_C(t)\|_2$, and the mean absolute Welch two-sample t-statistic across the p coordinates, $\mathrm{mean}_p\,|t_{\mathrm{Welch}}(t,p)|$. These two quantities identify the time indices at which the patient and control cohorts are most separable in $\boldsymbol{y}(t)$. Figure 3 supports the design claim that $\boldsymbol{W}^{(1)}(t)$ concentrates on the same temporal interval in which the geometry-module output $\boldsymbol{y}(t)$ shows the strongest cohort separation. Using the basis-corrected profile is important because clamped B-spline endpoints inflate $g(t)$ at $t \in \{0, T-1\}$ by a factor of approximately $1/0.68$ for purely geometric reasons; these endpoint spikes therefore do not reflect learned temporal content.

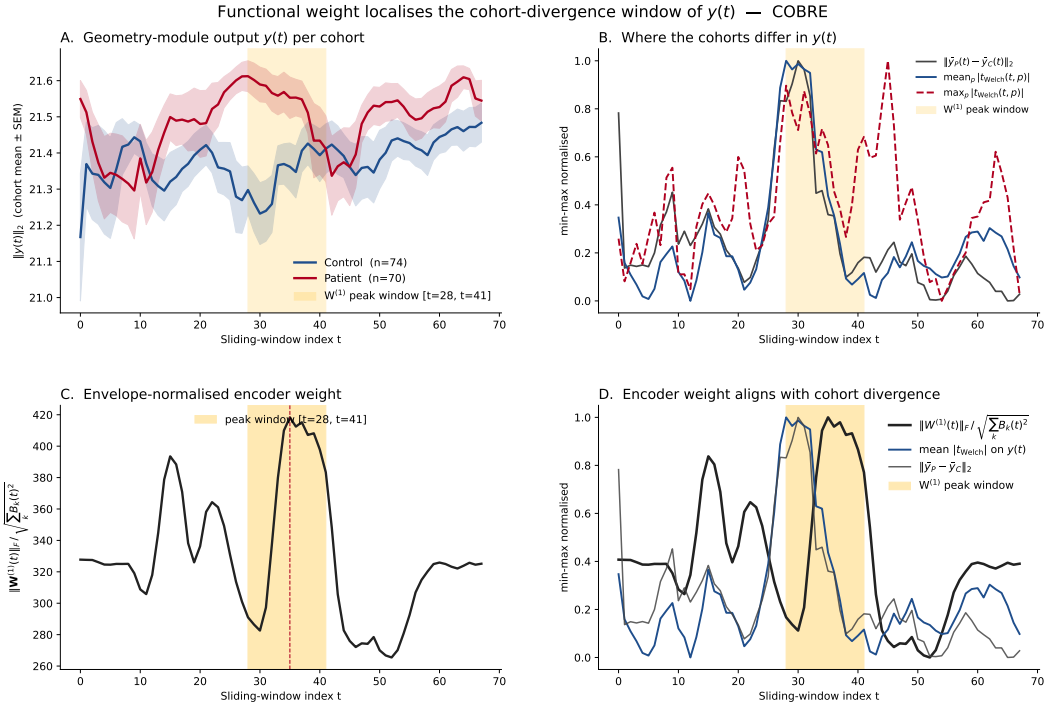

Figure 3: **The encoder functional weight localises the cohort-divergence window of $\boldsymbol{y}(t)$. (A)** Cohort-mean $\|\boldsymbol{y}(t)\|_2$ for healthy controls ($n = 74$) and schizophrenia patients ($n = 70$), with within-cohort SEM bands. **(B)** Two pointwise cohort-divergence summaries on $\boldsymbol{y}(t)$, namely $\|\bar{\boldsymbol{y}}_P(t) - \bar{\boldsymbol{y}}_C(t)\|_2$ and $\mathrm{mean}_p\,|t_{\mathrm{Welch}}(t,p)|$, both min-max normalised. **(C)** Basis-corrected encoder weight profile $\tilde{g}(t) = \|\boldsymbol{W}^{(1)}(t)\|_F / \sqrt{\sum_k B_k(t)^2}$, with the selected peak window shaded. **(D)** Min-max overlay of (B) and (C). The maximum of $\tilde{g}$ at $t = 35$ lies 5-7 windows away from the cohort-divergence maxima at $t = 28$ for $\mathrm{mean}_p\,|t_{\mathrm{Welch}}(t,p)|$ and $t = 30$ for $\|\bar{\boldsymbol{y}}_P(t) - \bar{\boldsymbol{y}}_C(t)\|_2$. All three peaks fall within the encoder peak window $[t_1, t_2] = [28, 41]$.

Connectivity contrast restricted to the peak window We next provide a static functional-connectivity visualisation of the contrast that `MatFAE` is expected to capture. Figure 4 shows the edges for which the schizophrenia cohort has higher mean Fisher-z correlation than the healthy-control cohort. This visualisation identifies the spatial structure of the static signal and helps explain why COBRE is the cohort on which `MatFAE` achieves its largest gain. To construct this static visualisation, each subject's SPD trajectory $\{\mathbf{X}_i(t)\}_{t=1}^{q}$, with $q = 68$ sliding-window covariance matrices and $m = 100$ Schaefer-100 7-network parcels, is collapsed to a single SPD matrix by the

Log-Euclidean Fréchet mean,

$$\bar{\mathbf{X}}_i \;=\; \mathrm{Exp}_{\mathbf{I}}^{\mathrm{LE}}\Big(\frac{1}{q}\sum_{t=1}^{q}\mathrm{Log}_{\mathbf{I}}^{\mathrm{LE}}(\mathbf{X}_i(t))\Big) \;=\; \exp\Big(\frac{1}{q}\sum_{t=1}^{q}\log\left(\mathbf{X}_i(t)\right)\Big). \tag{20}$$

The static covariance matrix $\bar{\mathbf{X}}_i$ is then converted to a correlation matrix $\mathbf{R}_i$ and Fisher-z transformed entry-wise. For each off-diagonal edge (j, k), a Welch two-sample t-statistic is computed between the patient ($n = 70$) and control ($n = 74$) cohorts using $\{\mathrm{atanh}([\mathbf{R}_i]_{j,k})\}_i$. The top 60 edges with the largest Welch $|t|$ in each direction are retained. Edges are displayed in panel A on a multi-view glass brain, using parcel centroids in MNI 2 mm space, and in panel B on a 100-node chord plot grouped by Yeo-7 network. Within each hemisphere, networks are ordered as Vis $\rightarrow$ SomMot $\rightarrow$ DorsAttn $\rightarrow$ SalVentAttn $\rightarrow$ Limbic $\rightarrow$ Cont $\rightarrow$ Default, so that both *Vis* arcs are placed at 12 o'clock and both *Default* arcs at 6 o'clock. The right hemisphere fills the right half of the circle clockwise from 12 o'clock, while the left hemisphere fills the left half counter-clockwise, mirroring the standard radiological axial view. Edge thickness and opacity scale linearly across the top-60 Welch $|t|$ range, with weaker edges drawn first so that the strongest edges appear on top.

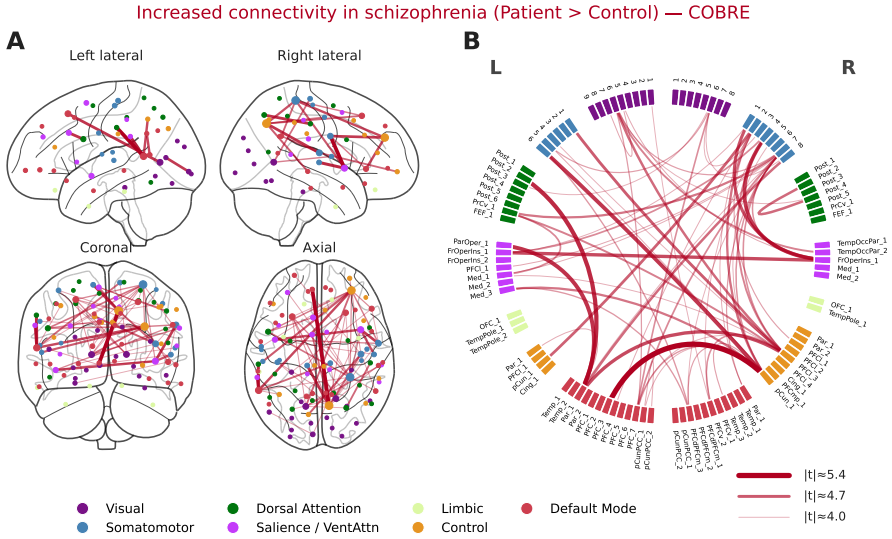

Figure 4: **Increased connectivity in schizophrenia on COBRE.** Top-60 edges by Welch $|t|$ where the schizophrenia cohort ($n = 70$) has higher Fisher-z correlation than the healthy-control cohort ($n = 74$). **(A)** Multi-view glass brain, including left lateral, right lateral, coronal, and axial views, with parcel centroids in MNI 2 mm space. Nodes are coloured by Yeo-7 network and sized by the total Welch $|t|$ at each node. **(B)** Chord plot with all 100 Schaefer-100 parcels on the rim, grouped by Yeo-7 network. The right hemisphere fills the right half of the circle clockwise from 12 o'clock, while the left hemisphere fills the left half counter-clockwise. Both Visual arcs are located at 12 o'clock and both Default-Mode arcs at 6 o'clock, with a cyclic gap at the top and an internal hemisphere gap at the bottom. Edge thickness and opacity scale linearly across the top-60 Welch $|t|$ range, with the weakest edges drawn first so that the strongest edges appear on top.

To test whether the encoder-identified peak window contains the strongest static functional-connectivity contrast, we recompute the LE-Fréchet mean in Eq. 20 using only $t \in [t_1, t_2] = [28, 41]$. We then repeat the Welch/Fisher-z edge-level test and retain the top 60 edges by $|t|$ for which the patient cohort exhibits higher correlation than the control cohort. Figure 5 shows that the canonical schizophrenia hyperconnectivity pattern is preserved when the average is restricted to the encoder-flagged window. The contrast remains dominated by connections among the Default-Mode, Salience/Ventral-Attention, and Frontoparietal-Control networks, with retained $|t| \in [3.9, 5.1]$, compared with $[4.0, 5.4]$ under all-window pooling. The modest reduction in the upper $|t|$ bound is consistent with the peak-window analysis using only $14/68 \approx 21\%$ of the temporal data, while still preserving the main spatial structure of the contrast.

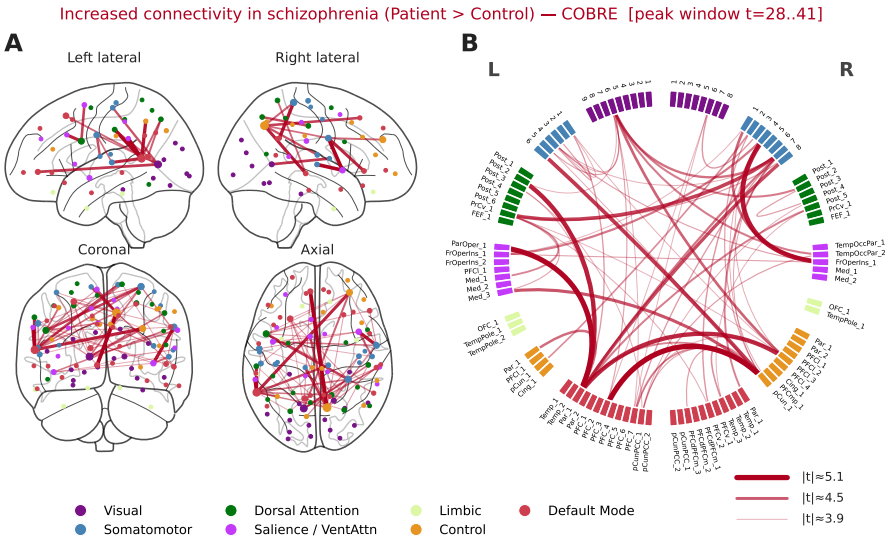

Figure 5: **Increased connectivity restricted to the encoder peak window.** Top-60 edges by Welch $|t|$ where the schizophrenia cohort has higher Fisher-z correlation than the healthy-control cohort, computed from per-subject LE-Fréchet means restricted to $t \in [28, 41]$, the window highlighted by $\boldsymbol{W}^{(1)}(t)$ in Figure 3. The same Default-Mode / Salience-Ventral-Attention / Frontoparietal-Control hyperconnectivity pattern is recovered, with retained $|t|$ values ranging from 3.9 to 5.1. This suggests that the encoder-highlighted window overlaps with the period in which the cohort-discriminative static signal is most concentrated.

Taken together, Figures 3-5 support the interpretability claim. The functional weight profile $\boldsymbol{W}^{(1)}(t)$ acts as an internally derived temporal saliency map, and its peak window aligns with the interval in which the underlying clinical contrast is most pronounced.

H LIMITATIONS

Individual-level diagnosis and patient/control classification from resting-state fMRI remain challenging because disease-related signals are often weak, distributed, and strongly confounded by scanner/site differences, preprocessing choices, parcellation schemes, and the unknown mental state of participants during scanning. Although we applied a common post-hoc processing pipeline whenever possible, the real datasets were not fully homogeneous in preprocessing or parcellation. Three cohorts used the Schaefer-100 atlas, whereas ABIDE-I used a 101-region Harvard-Oxford parcellation. Therefore, performance differences across cohorts may partly reflect differences in ROI definition, rather than only differences in disease phenotype or model behaviour.

Several additional sources of heterogeneity may affect the estimated dynamic functional connectivity trajectories. ADHD-200 and ABIDE-I are multi-site cohorts with TR values ranging from 1.5 to 3.0 s. We used site-specific band-pass filtering, but did not explicitly harmonise scanner, acquisition, or site effects; residual site-related variability may therefore remain. Upstream preprocessing also differed across consortia, including fMRIPrep for CNP, NIAK for COBRE, and CPAC for ADHD-200 and ABIDE-I. In particular, for ADHD-200 and ABIDE-I, only pre-extracted ROI time series were available, and the raw BOLD images and confound files were not redistributed. Consequently, retrospective Friston-24 motion regression and frame-wise scrubbing could not be applied, making nuisance-control procedures less directly comparable across datasets.

Finally, the two BiMap layers in the matrix-to-vector module contain $H(m_1 m + m_2 m_1) + 1$ trainable entries, independent of the time-series length. The geometry-aware decoder introduces two additional matrices, $\mathbf{M}_1$ and $\mathbf{M}_2$, containing $p_2 p_1 + m p_2$ trainable entries. All functional weights in FAEclust are represented using basis expansions, so only their expansion coefficients are trainable. The main parameter bottleneck is $\mathcal{W}^{(2)}(t)$ in the decoder. With K basis functions, the matrix

contains $O(Kp_2^4)$ trainable coefficients. This limits scalability to very high-dimensional parcellations, such as Schaefer-400 or Schaefer-1000, and motivates future work on more efficient geometric layers, low-rank approximations, or sparse SPD representations.