ACCELERATING BENCHMARKING OF FUNCTIONAL CONNECTIVITY MODELING VIA STRUCTURE-AWARE CORE-SET SELECTION

Ling Zhan[1,2], **Zhen Li**[1], **Junjie Huang**[1], **Tao Jia**[1,2,3*]

[1]College of Computer and Information Science, Southwest University, Chongqing, China
[2]Chongqing Key Laboratory of Brain-Inspired Cognitive Computing and
Educational Rehabilitation for Children with Special Needs, Chongqing Normal University, China
[3]College of Computer and Information Science, Chongqing Normal University, China
{zl0327, cs2003lz}@email.swu.edu.cn, {junjiehuang, tjia}@swu.edu.cn

ABSTRACT

Benchmarking the hundreds of functional connectivity (FC) modeling methods on large-scale fMRI datasets is critical for reproducible neuroscience. However, the combinatorial explosion of model–data pairings makes exhaustive evaluation computationally prohibitive, preventing such assessments from becoming a routine pre-analysis step. To break this bottleneck, we reframe the challenge of FC benchmarking by selecting a small, representative *core-set* whose sole purpose is to preserve the relative performance ranking of FC operators. We formalize this as a ranking-preserving subset selection problem and propose **S**tructure-aware **C**ontrastive **L**earning for **C**ore-set **S**election (**SCLCS**), a self-supervised framework to select these core-sets. **SCLCS** first uses an adaptive Transformer to learn each sample's unique FC structure. It then introduces a novel **S**tructural **P**erturbation **S**core (**SPS**) to quantify the stability of these learned structures during training, identifying samples that represent foundational connectivity archetypes. Finally, while **SCLCS** identifies stable samples via a top-k ranking, we further introduce a **density-balanced sampling strategy** as a necessary correction to promote diversity, ensuring the final core-set is both structurally robust and distributionally representative. On the large-scale REST-meta-MDD dataset, **SCLCS** preserves the ground-truth model ranking with just 10% of the data, outperforming state-of-the-art (SOTA) core-set selection methods by up to 23.2% in ranking consistency (nDCG@k). To our knowledge, this is the first work to formalize core-set selection for FC operator benchmarking, thereby making large-scale operators comparisons a feasible and integral part of computational neuroscience. Code is publicly available on https://github.com/lzhan94swu/SCLCS

1 INTRODUCTION

Methodological choices can substantially affect scientific reproducibility, as reflected in highly variable outcomes obtained from the same dataset, making systematic benchmarking increasingly important (Kohli et al., 2024; Qiu et al., 2024; Marek et al., 2022). This issue is especially acute in functional connectivity (FC) modeling, where hundreds of candidate statistical pairwise interactions (SPIs) require careful evaluation to ensure reliable conclusions (Liu et al., 2025; Roell et al., 2025). Yet the computational cost of exhaustive evaluation makes it impractical to run as a routine pre-analysis step for data-driven model selection (Ying et al., 2024; Zhou et al., 2021) (see the complexity analysis in **Appendix H**). To address this bottleneck, we propose a two-stage workflow: we first benchmark all candidate SPIs on a small, representative core-set to identify top performers, and then evaluate the selected SPI(s) on the full dataset for downstream analysis. This workflow hinges on selecting a core-set that preserves the relative performance ranking of SPIs.

[*]Corresponding author.

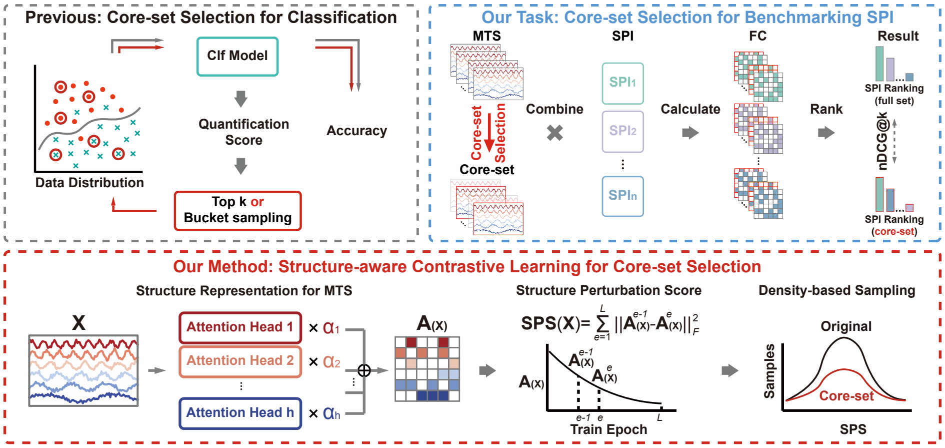

Figure 1: Overview of the **SCLCS** framework for ranking-preserving core-set selection. Contrasting with selection for single-model classification (top left), our task is to preserve the performance ranking of SPIs (top right). Our method (bottom) achieves this using a Transformer to learn structures, our novel **SPS** metric to ensure stability, and a density-aware strategy to promote diversity.

While core-set selection is well studied, most existing methods target a different goal: constructing a small training proxy for a single predictive model (Feldman, 2020; Lee et al., 2024; Hong et al., 2024b). In our setting (Figure 1), the core-set must preserve the relative performance ranking across hundreds of candidate SPIs (Liu et al., 2025; Cliff et al., 2023). This ranking-preservation objective raises three challenges: (1) Formulating a selection criterion that targets cross-SPI ranking stability rather than single-model training loss. (2) Defining a principled, structure-aware notion of sample importance based on FC patterns (the targets of SPIs). (3) Reducing the brittleness of score-based top-k selection, which can fail to generalize across sampling ratios and distort rankings.

In this work, we cast core-set selection for FC benchmarking as a ranking-preserving subset selection problem. Rather than training a predictive model, we seek a subset that preserves the SPI ordering from the full dataset (Figure 1). We evaluate on the REST-meta-MDD dataset (Yan et al., 2019; Long et al., 2020), a large multi-site resting-state fMRI dataset for MDD, which captures heterogeneity across acquisition sites and a large cohort. We instantiate benchmarking with two tasks, brain fingerprinting (Van De Ville et al., 2021) and MDD diagnosis (Gallo et al., 2023), both widely used in FC research (Lu et al., 2024; Otte et al., 2016). For each task, we score each SPI by how well the resulting FC matrices separate within-class from between-class pairs using Spearman's rank correlation (Sedgwick, 2014), yielding an SPI ranking. Core-set quality is measured by nDCG@k (Wang et al., 2013) between the SPI rankings induced by the core-set and the full dataset.

We use SPIs as a validation case because benchmarking FC operators has been formalized as a well-defined task in recent work (Liu et al., 2025; Cliff et al., 2023; Honari et al., 2021). Based on this formulation, we propose **S**tructure-aware **C**ontrastive **L**earning for **C**ore-set **S**election (**SCLCS**). As shown in Figure 1, **SCLCS** is built around a Transformer-based encoder that encodes sample-specific synchronization structure via an adaptively weighted fusion of attention heads. Under the assumptions of **Theorem 2**, we show this encoder has universal approximation capacity for continuous SPI mappings. We then define a **S**tructure **P**erturbation **S**core (**SPS**) to quantify the stability of these structures, and prioritize low-**SPS** samples to form a robust core-set. Because naïve top-k selection can be brittle for certain task structures, **SCLCS** augments it with a density-aware sampling strategy to improve diversity. **SCLCS** learns in an identity-supervised contrastive manner, using subject identities to encourage stable "brain fingerprints"(Van De Ville et al., 2021) that SPI-based analyses aim to capture(Liu et al., 2025; Luppi et al., 2024). This yields task-agnostic representations suitable for benchmarking. Finally, **SCLCS** is a pre-analysis acceleration tool that makes large-scale benchmarking computationally feasible, rather than a method for the final neuroscientific discovery task.

Our theoretical analysis and empirical results on 130 candidate SPIs support our design choices and show consistent improvements over strong baselines. The main contributions are: (1) We formulate core-set selection for efficient FC operator (SPI) benchmarking as a ranking-preservation problem. (2) We propose **SCLCS**, a structure-aware framework for selecting stable and diverse samples for ranking-based benchmarking. (3) We provide a universal approximation result for continuous SPI mappings (Theorem 2) and introduce **SPS**, a new use of attention dynamics to quantify structural heterogeneity. (4) We show that **SCLCS** enables reliable benchmarking at a fraction of the computational cost, making large-scale comparisons practical.

2 RELATED WORK

Benchmarking in Functional Connectivity. Selecting an appropriate SPI (i.e., a network modeling method) is a central challenge in modern FC research. Prior studies show that different SPIs can yield divergent FC topologies and, consequently, different scientific conclusions (Smith et al., 2011; 2013; Bobadilla-Suarez et al., 2020; Mohanty et al., 2020; Honari et al., 2021; Luppi et al., 2024), contributing to long-standing concerns about reproducibility (Collaboration, 2015; Botvinik-Nezer et al., 2020; Marek et al., 2022). Meanwhile, comprehensive libraries such as `pyspi`(Cliff et al., 2023), which include hundreds of SPIs, highlight the methodological richness of the field and magnify the scale of the selection problem(Liu et al., 2025; Roell et al., 2025). These works motivate systematic benchmarking, but the computational cost of evaluating large SPI suites remains a key practical bottleneck. We address this bottleneck by introducing a core-set selection approach for efficient SPI benchmarking.

Core-set Selection. Core-set selection is a fundamental problem in machine learning (Guo et al., 2022; Feldman, 2020; Ros & Guillaume, 2019). Most score-based (Coleman et al., 2020; Feldman & Zhang, 2020; Paul et al., 2021; Toneva et al., 2019) and diversity-based (Sener & Savarese, 2018; Xia et al., 2022) methods construct proxy datasets for training a single predictive model, which mismatches our ranking-preservation objective over many SPIs. Their criteria are often model-dependent (e.g., EVA (Hong et al., 2024b)) and typically assume static i.i.d. inputs, overlooking the temporal dependencies in fMRI time series from which FC structure is derived. Consequently, selected core-sets may not transfer to ranking-based evaluation over large SPI suites (Liu et al., 2024; Lee et al., 2024). Training-acceleration methods (Hong et al., 2024a; Killamsetty et al., 2021; Mirzasoleiman et al., 2020; Wei et al., 2015) share the same single-model focus and thus do not directly address our setting.

To our knowledge, **SCLCS** is among the first methods tailored for accelerating benchmarking of FC operators (SPIs) via core-set selection. It uses the stability of learned synchronization structures during training as a selection criterion, which is particularly natural for neuroimaging. While related to graph structure learning (Li et al., 2023; Zhou et al., 2023; Zong et al., 2024), **SCLCS** treats learned structure as a diagnostic probe rather than an inference output. We therefore do not review general-purpose structure learning in depth.

3 PRELIMINARIES

Benchmarking FC modeling. The goal of FC benchmarking is to produce a principled ranking of statistical pairwise interaction (SPI) operators from a set $\mathcal{S}$. For a given fMRI dataset $\mathcal{X}$, where each sample $\mathbf{X} \in \mathcal{X}$ is a matrix in $\mathbb{R}^{N \times T}$, each operator $S \in \mathcal{S}$ maps $\mathbf{X}$ to an FC matrix $S(\mathbf{X}) \in \mathbb{R}^{N \times N}$. An evaluation index, $\mathcal{I} : \mathcal{S} \to \mathbb{R}$, assigns a score to each SPI under a chosen evaluation protocol. This process induces a ranking over all SPIs in $\mathcal{S}$, which we denote as $\text{Rank}(\mathcal{S}, \mathcal{X})$, to guide the selection of a suitable SPI for subsequent analysis.

Core-set Selection for Benchmarking FC Modeling. Computing the full-dataset ranking $\text{Rank}(\mathcal{S}, \mathcal{X})$ is often computationally prohibitive. We therefore introduce the task of core-set selection for benchmarking, which seeks to identify a small subset of samples $\mathcal{X}_c \subset \mathcal{X}$ where $|\mathcal{X}_c| \ll |\mathcal{X}|$ that acts as an efficient proxy. Formally, a high-quality core-set is the solution to the following optimization problem:

$$\mathcal{X}_c^* = \underset{\mathcal{X}' \subset \mathcal{X},\, |\mathcal{X}'| = c}{\text{argmin}} \; \mathcal{D}\big(\text{Rank}(\mathcal{S}, \mathcal{X}),\, \text{Rank}(\mathcal{S}, \mathcal{X}')\big). \tag{1}$$

where $\mathcal{D}(\cdot, \cdot)$ is a ranking discrepancy metric (e.g., based on nDCG@k) and c is the core-set budget. The goal is to preserve the full-dataset ranking while using only c samples.

Directly optimizing this objective is intractable, as it requires exhaustively evaluating an exponential number of subsets, with each evaluation incurring the very computational cost we aim to avoid. We therefore propose a practical proxy: selecting a structurally representative subset. The core hypothesis is that preserving the distribution of functional connectivity structures also preserves the SPI ranking. Our **SCLCS** framework, detailed next, is designed to find such a subset.

4 Method

In this section, we introduce the detailed formulation of the proposed **SCLCS**. **SCLCS** consists of four modules: (1) attention-based FC learning, (2) structural perturbation score calculation, (3) structure-aware density-balanced sampling, and (4) contrastive learning.

4.1 Attention-based FC Learning

To select a rank-preserving core-set, our framework first requires an encoder that can learn a general and expressive representation of each sample's FC structure. The self-attention mechanism within Transformers is a natural candidate for this task, as it can model complex inter-regional relationships (Vaswani et al., 2017). However, naïve fusion of multiple attention heads via uniform averaging is insufficient, as it can obscure distinct structural patterns learned by individual heads **Theorem 1** (proved in **Appendix A**).

Theorem 1 (Interference of Averaged Attention)**.** *Let $\{\mathbf{A}_h\}_{h=1}^H$ be row-stochastic attention matrices. Assume disjoint structural masks: for each row i there exist pairwise-disjoint sets $\{S_h^{(i)}\}_{h=1}^H$ such that $\mathbf{A}_h(i,j) = 0$ for all $j \notin S_h^{(i)}$. Let $\bar{\mathbf{A}} := \frac{1}{H}\sum_{h=1}^H \mathbf{A}_h$. Then for every row i:*

$$\mathrm{supp}\big(\bar{\mathbf{a}}^{(i)}\big) = \bigcup_{h=1}^{H} S_h^{(i)} \quad \textit{and} \quad \mathcal{H}\big(\bar{\mathbf{a}}^{(i)}\big) > \min_{1\le h\le H} \mathcal{H}\big(\mathbf{a}_h^{(i)}\big) \ \textit{if}\ \{\mathbf{a}_h^{(i)}\}_{h=1}^H \ \textit{are not all identical.}$$

In particular, if $H \ge 2$, naive averaging expands support beyond any single head's mask and inflates entropy, blurring head-specific structure.

Theorem 1 pertains strictly to the internal attention maps and explains the empirical failure of directly applying the traditional Transformer on core-set selection for benchmarking FC modeling. It motivates us to propose an adaptive fusion mechanism that aggregates head-specific attention matrices via learnable weights. This modification is not merely an engineering choice: we prove it endows the architecture with the power of a universal approximator for the class of continuous FC operators as formalized in **Theorem 2** (proved in **Appendix B**). This provides a theoretical foundation for its ability to capture the diverse synchronization patterns required for our benchmarking task.

Theorem 2 (Universal Approximation of Continuous Stochastic SPIs[1])**.** *Let $\mathcal{X} \subset \mathbb{R}^{N\times T}$ be compact. Let $S : \mathcal{X} \to \Delta^{N-1\times N}$ be continuous, where $\Delta^{N-1\times N} := \{\mathbf{P} \in \mathbb{R}^{N\times N} : \mathbf{P}\mathbf{1} = \mathbf{1},\ \mathbf{P} \ge 0\}$ denotes the set of row-stochastic matrices. Consider the adaptive multi-head attention family*

$$\mathbf{A}_\theta(\mathbf{X}) \ := \ \sum_{h=1}^{H} \alpha_h \ \mathrm{softmax}\Big(\frac{\mathbf{X}\mathbf{W}_h^Q(\mathbf{W}_h^K)^\top\mathbf{X}^\top}{\tau}\Big), \qquad \boldsymbol{\alpha} \in \Delta^{H-1},\ \tau > 0, \tag{2}$$

where softmax *is applied row-wise. Then for every $\varepsilon > 0$ there exist H, τ, and parameters $\{\mathbf{W}_h^Q, \mathbf{W}_h^K\}_{h=1}^H$ and $\boldsymbol{\alpha}$ such that*

$$\sup_{\mathbf{X}\in\mathcal{X}} \big\|\mathbf{A}_\theta(\mathbf{X}) - S(\mathbf{X})\big\|_F < \varepsilon. \tag{3}$$

Our implementation is as follows: for each fMRI sample $\mathbf{X} \in \mathbb{R}^{N\times T}$, we treat the N ROIs as input tokens, where each token has a feature dimension of T. Each attention head independently projects queries and keys using learnable linear maps, parameterized by matrices $\mathbf{W}_h^Q$ and $\mathbf{W}_h^K$:

$$\mathbf{Q}_h = \mathbf{X}\mathbf{W}_h^Q, \quad \mathbf{K}_h = \mathbf{X}\mathbf{W}_h^K, \quad \mathbf{W}_h^Q, \mathbf{W}_h^K \in \mathbb{R}^{D\times d}, \quad h = 1, \ldots, H. \tag{4}$$

[1] This statement is for continuous targets on compact domains. If an SPI uses discrete thresholds (hard masks), the guarantee applies to any continuous relaxation (e.g., finite-temperature softmax / sigmoid gates) and then a limiting argument is required to justify the hard-threshold limit.

The attention matrix from head h is computed as:

$$\mathbf{A}_h = \text{softmax}\left(\frac{\mathbf{Q}_h\mathbf{K}_h^\top}{\sqrt{d}}\right), \quad \mathbf{A}_h \in \mathbb{R}^{N\times N}. \tag{5}$$

Motivated by **Theorem 1**, uniform head averaging is a structural constraint: $\alpha_h \equiv 1/H$ forces every sample to use the same (high-entropy) centroid of head-wise attention patterns. By strict concavity of Shannon entropy, this averaging inflates uncertainty and smears head-specific structure. Learnable fusion weights $\boldsymbol{\alpha}$ relax the constraint, enabling sparse/peaked mixtures (up to single-head selection) to reduce interference while still combining complementary patterns. Thus the operator class strictly expands: by **Theorem 2**, an adaptive fusion module can approximate continuous FC operator on compact domains.

Thus, we propose a learnable fusion mechanism that aggregates head-specific attention matrices via adaptive weights, formulated as:

$$\mathbf{A} = \sum_{h=1}^{H} \boldsymbol{\alpha}_h \mathbf{A}_h, \quad \text{with} \quad \sum_{h=1}^{H} \boldsymbol{\alpha}_h = 1, \quad \boldsymbol{\alpha}_h \geq 0, \tag{6}$$

where the weight $\boldsymbol{\alpha}$ is normalized via `softmax`. The resulting matrix $\mathbf{A} \in \mathbb{R}^{N\times N}$ serves as the operational definition of a sample's FC structure, forming the basis for our selection criteria. Importantly, we treat these attention maps as a normalized structural probe—a sample-specific proxy for synchronization structure—not as an attempt to replicate the raw outputs of any particular SPI.

4.2 Structural Perturbation Score (SPS)

Having established an encoder that is expressive enough to capture diverse FC structures, the next challenge is to define a criterion for identifying the most fundamental samples for a robust benchmark. Our central hypothesis is that samples representing common, foundational connectivity patterns will induce stable structural representations during training, while noisy or atypical samples will cause greater fluctuations. To quantify this phenomenon, we propose the **S**tructural **P**erturbation **S**core (**SPS**), a metric grounded in the principle that perturbation magnitude reflects structural heterogeneity (**Proposition 1**).

Proposition 1 (Mixture-driven perturbation magnitude). *Let $S^{(1)}, \dots, S^{(K)} \in \mathbb{R}^{N\times N}$ be distinct prototypes and $D_{kl} := \|S^{(k)} - S^{(l)}\|_F^2$ ($D_{kl} > 0$ for $k \neq l$). Let $(Z_e)_{e\geq 1}$ be i.i.d. with $\Pr[Z_e = S^{(k)}] = \lambda_k$, $\sum_k \lambda_k = 1$, and define $\Delta_e := \|Z_e - Z_{e-1}\|_F^2$.*

Then

$$\mathbb{E}[\Delta_e] = \sum_{k,l} \lambda_k \lambda_l D_{kl} = 2\sum_{k<l} \lambda_k \lambda_l D_{kl}. \tag{7}$$

With $D_{\min} := \min_{k<l} D_{kl}$ and $D_{\max} := \max_{k<l} D_{kl}$,

$$D_{\min}\Big(1 - \sum_k \lambda_k^2\Big) \leq \mathbb{E}[\Delta_e] \leq D_{\max}\Big(1 - \sum_k \lambda_k^2\Big). \tag{8}$$

In particular, $\mathbb{E}[\Delta_e]$ scales with the Gini impurity $1 - \sum_k \lambda_k^2$ up to constants set by prototype separation. If $D_{kl} \equiv D$ for all $k \neq l$, then

$$\mathbb{E}[\Delta_e] = D\Big(1 - \sum_k \lambda_k^2\Big). \tag{9}$$

(Proof in Appendix C.)

Proposition 1 indicates samples that are a purer representation of a single archetype will be more stable (low **SPS**). The **SPS** for a sample $\mathbf{X} \in \mathcal{X}$ is thus defined as the cumulative structural instability across L training epochs:

$$\text{SPS}(\mathbf{X}) = \frac{1}{L}\sum_{e=1}^{L} \left\| \mathbf{A}_{(\mathbf{X})}^{(e)} - \mathbf{A}_{(\mathbf{X})}^{(e-1)} \right\|_F^2, \tag{10}$$

where $\mathbf{A}^{(e)}_{(\mathbf{X})}$ denotes the attention-based structure matrix of sample $\mathbf{X} \in \mathcal{X}$ at training epoch $e \in \{1, \ldots, L\}$ and $|| \cdot ||_F$ denotes the Frobenius norm. Specifically, $\mathbf{A}^{(e)}_{(\mathbf{X})}$ is a proxy for structural representation, not a proposed network model. In this way, **SPS** captures the structural volatility of the sample-specific synchronization graph during training, not fidelity to any specific SPI. Our rationale is that a robust and reliable benchmark is built upon foundational, structurally stable samples. As supported by **Proposition 1**, low-**SPS** samples exhibit less internal structural conflict and thus represent stable archetypes of functional connectivity. Therefore, our primary selection strategy is to rank samples by their **SPS** and select those with the lowest **SPS**.

For **SPS** to be a reliable metric, however, we must ensure that it is a consistent estimator that does not depend on the arbitrary length of the training process. **Lemma 1** (in **Appendix D**) provides this theoretical guarantee: Under stable perturbation dynamics, the assumptions of **Lemma 1** are satisfied. We used extensive grid search to find a configuration that achieves optimal performance on the downstream ranking preservation task as detailed in **Appendix M**. Notably, the assumptions of that configuration's stationarity and ergodicity in **Lemma 1** are empirically supported by our convergence analysis in **Appendix K.2**, where we show that the perturbation dynamics stabilize as the model converges. This standard procedure sufficiently validates the feasibility of **SPS**.

4.3 Structure-aware Density-Balanced Sampling

While selecting for structurally stable (low-**SPS**) samples provides a robust foundation, a naïve top-k selection risks creating a core-set with low diversity by over-selecting from dense clusters of typical patterns. This lack of diversity can cause the core-set benchmark to diverge from the full-dataset ranking (**Theorem 3**).

Theorem 3 (Persistent bias of top-k selection)**.** *Let $\mathcal{X}$ contain two clusters C_p, C_q with proportions π_p, π_q. Given $\mathbf{x} \in C_r$ ($r \in \{p, q\}$), let the score $s(\mathbf{x})$ have continuous CDF F_r, and assume scores are independent across samples. Select the $k = \lfloor \rho N \rfloor$ samples with the smallest scores ($\rho \in (0,1)$), and write $\widehat{\pi}_r := |S_k \cap C_r|/k$.*

Let τ satisfy the mixture-quantile equation

$$\pi_p F_p(\tau) + \pi_q F_q(\tau) = \rho, \tag{11}$$

and assume strict separation at τ:

$$\gamma := F_p(\tau) - F_q(\tau) > 0. \tag{12}$$

Then

$$\widehat{\pi}_p \xrightarrow{\Pr} \frac{\pi_p F_p(\tau)}{\rho} = \pi_p + \delta, \qquad \widehat{\pi}_q \xrightarrow{\Pr} \frac{\pi_q F_q(\tau)}{\rho} = \pi_q - \delta, \qquad \delta := \frac{\pi_p \pi_q}{\rho} \gamma > 0. \tag{13}$$

Consequently, the representation error $\Delta_k := |\widehat{\pi}_p - \pi_p| + |\widehat{\pi}_q - \pi_q|$ satisfies $\Delta_k \xrightarrow{\Pr} 2\delta > 0$. (See **Appendix E**.)

To explicitly balance stability with diversity, we introduce a density-aware sampling scheme, yielding the **SCLCS**$_{\text{Dense}}$ variant. This scheme first ensures robustness by retaining a pool of the most stable samples (the bottom $1 - \beta$ quantile of **SPS** scores), then promotes diversity by applying Kernel Density Estimation (KDE) (Węglarczyk, 2018) to up-weight samples from sparser regions within that stable pool. This mitigates redundancy and ensures the core-set captures a broader range of structurally distinct subtypes, which is crucial for including less common but potentially critical neural patterns often associated with clinical biomarkers.

Specifically, given the set of **SPS** for all samples in $\mathcal{X}$, we first discard the top β quantile of the most unstable samples to form a stable candidate pool $\tilde{\mathcal{X}}$. This is formally defined as:

$$\tilde{\mathcal{X}} = \{\mathbf{X} \in \mathcal{X} \mid \mathbf{SPS}(\mathbf{X}) \leq Q_{1-\beta}\}, \tag{14}$$

where $Q_{1-\beta}$ is the empirical quantile. On $\tilde{\mathcal{X}}$, we fit a Gaussian KDE to the empirical distribution of $\{\mathrm{SPS}(\mathbf{X}) : \mathbf{X} \in \tilde{\mathcal{X}}\}$, and define the local density of a sample as the KDE evaluated at its **SPS**:

$$\hat{p}_{\mathrm{SPS}}(s) = \mathtt{KDE}\big(\{\mathrm{SPS}(\mathbf{X})\}_{\mathbf{X} \in \tilde{\mathcal{X}}}\big), \qquad \rho(\mathbf{X}) = \hat{p}_{\mathrm{SPS}}\big(\mathrm{SPS}(\mathbf{X})\big). \tag{15}$$

To promote diversity, weights are set inversely proportional to $\rho(\mathbf{X})$ and normalized over $\tilde{\mathcal{X}}$:

$$w(\mathbf{X}) = \frac{1}{\rho(\mathbf{X}) + \epsilon}, \qquad w(\mathbf{X}) \leftarrow \frac{w(\mathbf{X})}{\sum_{\mathbf{X}' \in \tilde{\mathcal{X}}} w(\mathbf{X}')}. \tag{16}$$

We then select m samples without replacement from $\tilde{\mathcal{X}}$ using weights $\{w(\mathbf{X})\}$. This yields a structurally diverse subset that up-weights samples in low-density regions of the **SPS** distribution within the stable pool. Theoretical guarantees on coverage and benchmarking consistency are demonstrated in **Theorem 4** and **Theorem 5**, respectively (provided and proved in **Appendix F** and **Appendix G**, respectively). Notably, this scheme is applicable to other score-based methods by replacing **SPS** to different metrics as shown in **Appendix J**.

4.4 Structure-aware Contrastive Learning

To learn structural representations, we train the encoder with a structure-aware contrastive objective. Motivated by FC evaluation practices that exploit inter-subject differences (Liu et al., 2025; Luppi et al., 2024), we enforce consistency among samples drawn from the same subject within a scan session. This identity-supervised setup encourages the model to capture stable, person-specific traits ("brain fingerprints" (Lu et al., 2024)), providing a task-agnostic signal for structure-based selection.

First, to obtain a graph-level embedding for each sample, we compute node-level embeddings $\mathbf{Z} \in \mathbb{R}^{N \times d}$ by applying the learned attention matrix to the value embeddings and projecting the result through a final linear layer. A global mean pooling is then applied to obtain the graph-level embedding $\mathbf{z} \in \mathbb{R}^d$, which captures the sample's global topological semantics and serves as input to our contrastive loss:

$$\mathcal{L}_{\text{contrast}} = \frac{1}{|\mathcal{P}|} \sum_{(i,j) \in \mathcal{P}} -\log \frac{\exp\left(\text{sim}(\mathbf{z}_i, \mathbf{z}_j)/\tau\right)}{\sum_{k \in \mathcal{N}(i)} \exp\left(\text{sim}(\mathbf{z}_i, \mathbf{z}_k)/\tau\right)}, \tag{17}$$

where sim$(\cdot, \cdot)$ is cosine similarity, τ is a temperature parameter, and the pairs are defined across a batch. Positive pairs $(i, j) \in \mathcal{P}$ consist of two different temporal segments from the same subject, while for a given anchor sample i, the negative sample $k \in \mathcal{N}(i)$ are all other samples in the batch from different subjects. All the trainable parameters are optimized using Adam (Kingma & Ba, 2015).

5 Experiment

In this section, we present empirical results to validate our proposed framework. We first detail the experimental settings and then present the main quantitative and qualitative results comparing **SCLCS** to SOTA baselines. Due to space constraints, several important settings and supplementary analyses, including detailed experimental settings for reproduction (**Appendix M**), a detailed computational cost breakdown (**Appendix H**), the effect of supervised information on baselines (**Appendix I**), the application of our density-aware sampling to other baselines (**Appendix J**), empirical results (**Appendix K**) for supporting **Theorem 2** and the assumptions in **Lemma 1**, and additional generalization and robustness analysis (**Appendix L**).

5.1 Experimental Settings

Data We validate our framework on the REST-meta-MDD dataset (Yan et al., 2019), a large-scale, multi-site resting-state fMRI collection comprising 1,642 subjects from 17 sites. This highly heterogeneous collection was released with a standardized preprocessing pipeline, providing a rigorous testbed for core-set selection. For efficiency and consistency with prior work, we focus on a subset of 904 subjects (Long et al., 2020). To capture dynamic patterns, each subject's fMRI record is segmented into overlapping temporal samples via a sliding window, yielding 4,520 samples. Demographic statistics are in Table 1. Complete data and preprocessing details are in **Appendix M.1**.

Baselines Following the experimental setup in Hong et al. (2024b), we extend their comparison with two additions (k-Means and BOSS), evaluating against 9 baselines in total: (1) Random; (2) k-Means (Hartigan & Wong, 1979); (3) Forgetting score (Toneva et al., 2019); (4) Entropy (Coleman

Table 1: Summary of the used subset of REST-meta-MDD.

Site	#Samples	#HC	#MDD	#Male	#Female	Age Range	Education Range (Years)
15	335	37	30	26	41	19–65	5–21
17	410	41	41	27	55	18–30	9–17
19	245	31	18	19	30	18–51	5–15
20	2395	229	250	157	322	18–65	3–20
21	720	65	79	62	82	18–65	5–15
22	190	20	18	21	17	19–47	8–17
23	225	23	22	18	27	19–54	6–20
Overall	4520	458	446	330	574	18–65	3–21

et al., 2020); (5) EL2N (Paul et al., 2021); (6) AUM (Pleiss et al., 2020); (7) CCS (Zheng et al., 2023); (8)EVA (Hong et al., 2024b);(9) BOSS (Acharya et al., 2024). The latter 7 of them are SOTA methods designed for core-set selection. Detailed introduction is summarized in **Appendix M.2**.

Environment Experiments are performed on an 8-GPU (H20) high-performance computing cluster provided by the Large-scale Instrument Sharing Platform of Southwest University.

Evaluation Protocol Our goal is to select a subset that preserves model ranking, not to train a single predictive model. Therefore, our evaluation deviates from the standard train/test split. The protocol is as follows: (1) Each method selects a core-set of a given size from the entire dataset. (2) We then compute the SPI performance ranking on both the full dataset (ground truth) and the selected core-set. (3) The quality of the core-set is measured by the consistency between these two rankings.

Task We use two distinct downstream tasks to evaluate SPI discriminability: brain fingerprinting (distinguishing individuals based on subject ID), which probes for fine-grained, subject-specific structures, and MDD diagnosis, which relies on cohort-level patterns.

Metrics We use two primary metrics. (1) **Discriminability Score**, a metric based on Spearman's rank correlation (Sedgwick, 2014) that quantifies the class separability (within- vs. between-class) of the resulting FC matrices. (2) **Ranking Consistency**, the concordance between the full-dataset and core-set SPI rankings using nDCG@$5/10/20$ (Wang et al., 2013).

5.2 Quantitative Results

We present the primary quantitative results in Table 2 and Table 3. The evaluation includes **SCLCS** (using low-**SPS** top-k selection), **SCLCS**$_{\text{Dense}}$ (density-aware selection), and **SPS**$_{\text{MHA}}$ (a variant using naïve attention averaging to empirically test **Theorem 1**).

Table 2: Performance of different methods on brain fingerprinting ranking task (mean ± std) and nDCG@k is reported as percentage (×100).

Method	nDCG@5			nDCG@10			nDCG@20		
	0.1	0.3	0.5	0.1	0.3	0.5	0.1	0.3	0.5
Random	$15.17_{\pm 13.97}$	$69.57_{\pm 24.69}$	$66.60_{\pm 20.21}$	$17.97_{\pm 17.50}$	$71.19_{\pm 20.93}$	$66.98_{\pm 19.47}$	$21.97_{\pm 21.91}$	$71.63_{\pm 21.27}$	$68.13_{\pm 17.80}$
k-Means	$17.32_{\pm 11.53}$	$65.72_{\pm 12.25}$	$67.29_{\pm 11.41}$	$21.23_{\pm 15.32}$	$64.35_{\pm 14.83}$	$63.37_{\pm 9.92}$	$22.15_{\pm 15.19}$	$62.58_{\pm 19.22}$	$66.78_{\pm 15.28}$
Forgetting	$14.41_{\pm 7.84}$	$54.43_{\pm 13.26}$	$43.36_{\pm 3.35}$	$15.49_{\pm 7.83}$	$48.57_{\pm 7.85}$	$49.87_{\pm 4.66}$	$20.23_{\pm 7.21}$	$55.80_{\pm 6.72}$	$49.70_{\pm 4.60}$
Entropy	$47.40_{\pm 40.26}$	$22.84_{\pm 13.20}$	$59.95_{\pm 26.99}$	$37.73_{\pm 25.94}$	$32.05_{\pm 15.22}$	$58.68_{\pm 23.71}$	$36.05_{\pm 18.11}$	$35.90_{\pm 16.19}$	$57.72_{\pm 21.84}$
El2N	$35.56_{\pm 36.16}$	$20.30_{\pm 5.77}$	$31.03_{\pm 34.57}$	$36.51_{\pm 25.38}$	$33.18_{\pm 14.68}$	$32.70_{\pm 30.84}$	$33.30_{\pm 21.55}$	$35.82_{\pm 12.10}$	$40.06_{\pm 32.08}$
AUM	$65.92_{\pm 33.80}$	$56.68_{\pm 11.11}$	$38.17_{\pm 13.94}$	$60.95_{\pm 30.91}$	$62.05_{\pm 4.91}$	$36.83_{\pm 8.78}$	$51.75_{\pm 22.42}$	$59.09_{\pm 5.72}$	$38.34_{\pm 6.88}$
CCS	$1.90_{\pm 2.07}$	$30.53_{\pm 14.15}$	$46.65_{\pm 22.32}$	$2.92_{\pm 3.24}$	$29.13_{\pm 8.00}$	$51.78_{\pm 24.12}$	$16.24_{\pm 13.34}$	$32.56_{\pm 14.15}$	$52.18_{\pm 20.20}$
EVA	$38.40_{\pm 40.57}$	$62.03_{\pm 26.64}$	$37.80_{\pm 42.45}$	$43.37_{\pm 19.92}$	$55.01_{\pm 20.05}$	$49.56_{\pm 33.28}$	$43.22_{\pm 15.28}$	$53.51_{\pm 14.70}$	$65.49_{\pm 21.99}$
BOSS	$15.98_{\pm 25.58}$	$42.11_{\pm 24.33}$	$35.36_{\pm 9.21}$	$29.44_{\pm 11.05}$	$40.57_{\pm 23.37}$	$39.15_{\pm 6.71}$	$31.45_{\pm 9.45}$	$38.24_{\pm 19.65}$	$38.92_{\pm 5.97}$
SCLCS	$\mathbf{81.21}_{\pm 2.86}$	$50.24_{\pm 13.16}$	$\mathbf{72.68}_{\pm 20.83}$	$\mathbf{66.54}_{\pm 1.10}$	$49.45_{\pm 14.18}$	$\mathbf{71.86}_{\pm 4.35}$	$\mathbf{57.46}_{\pm 0.52}$	$53.40_{\pm 16.27}$	$\mathbf{70.13}_{\pm 1.57}$
SCLCS$_{\text{Dense}}$	$35.73_{\pm 31.18}$	$\mathbf{79.18}_{\pm 6.29}$	$51.54_{\pm 13.83}$	$35.84_{\pm 28.43}$	$\mathbf{73.45}_{\pm 1.42}$	$50.54_{\pm 15.16}$	$41.03_{\pm 35.89}$	$\mathbf{72.96}_{\pm 3.35}$	$55.43_{\pm 13.64}$
SPS$_{\text{MHA}}$	$1.32_{\pm 2.07}$	$12.23_{\pm 5.11}$	$15.62_{\pm 6.33}$	$2.92_{\pm 1.13}$	$13.12_{\pm 11.33}$	$11.28_{\pm 4.32}$	$1.21_{\pm 1.04}$	$12.13_{\pm 3.17}$	$12.18_{\pm 7.23}$

Brain Fingerprinting This task rewards subject-specific patterns, which aligns with our identity-supervised objective. As shown in Table 2, **SCLCS** achieves stronger performance with lower variance at sampling ratios 0.1 and 0.5, suggesting that selecting structurally stable samples via low-**SPS** top-k ranking is effective. In contrast, **SPS**$_{\text{MHA}}$ performs poorly, consistent with **Theorem 1**. At the moderate ratio 0.3, **SCLCS** degrades, whereas **SCLCS**$_{\text{Dense}}$ performs best. This pattern supports **Theorem 3**: when stability ranking alone is insufficient, explicitly promoting structural diversity provides a corrective signal.

Table 3: Performance of different methods on MDD diagnosis ranking task (mean ± std) and nDCG@k is reported as percentage (×100).

Method	**nDCG@5**			**nDCG@10**			**nDCG@20**		
	0.1	0.3	0.5	0.1	0.3	0.5	0.1	0.3	0.5
Random	$49.58_{\pm 33.78}$	$23.77_{\pm 35.85}$	$31.76_{\pm 15.83}$	$\underline{60.50}_{\pm 33.36}$	$25.81_{\pm 36.19}$	$77.95_{\pm 11.77}$	$67.04_{\pm 27.03}$	$30.30_{\pm 39.67}$	$82.71_{\pm 8.59}$
k-Means	$51.32_{\pm 17.47}$	$32.45_{\pm 33.46}$	$37.33_{\pm 17.20}$	$30.18_{\pm 12.65}$	$40.47_{\pm 14.81}$	$43.75_{\pm 22.84}$	$28.62_{\pm 14.33}$	$37.87_{\pm 14.77}$	$79.83_{\pm 5.24}$
Forgetting	$28.70_{\pm 28.06}$	$41.60_{\pm 28.08}$	$61.42_{\pm 4.31}$	$40.56_{\pm 28.06}$	$50.27_{\pm 32.37}$	$66.72_{\pm 8.60}$	$44.37_{\pm 28.07}$	$57.02_{\pm 33.39}$	$75.70_{\pm 7.58}$
Entropy	$29.00_{\pm 45.10}$	$31.84_{\pm 46.19}$	$57.48_{\pm 29.81}$	$29.17_{\pm 43.24}$	$40.75_{\pm 44.05}$	$63.23_{\pm 23.44}$	$30.30_{\pm 43.52}$	$45.85_{\pm 45.00}$	$70.79_{\pm 18.00}$
El2N	$30.93_{\pm 41.17}$	$51.70_{\pm 28.14}$	$68.85_{\pm 12.46}$	$36.05_{\pm 24.41}$	$58.54_{\pm 27.55}$	$72.96_{\pm 15.77}$	$41.04_{\pm 39.61}$	$64.68_{\pm 26.25}$	$78.85_{\pm 14.56}$
AUM	$36.07_{\pm 12.59}$	$42.68_{\pm 46.16}$	$35.49_{\pm 14.37}$	$39.17_{\pm 14.58}$	$44.41_{\pm 44.05}$	$58.14_{\pm 16.54}$	$44.94_{\pm 18.48}$	$48.44_{\pm 42.48}$	$64.36_{\pm 14.40}$
CCS	$\underline{55.95}_{\pm 16.28}$	$59.92_{\pm 30.08}$	$74.73_{\pm 12.46}$	$59.25_{\pm 16.26}$	$67.01_{\pm 24.61}$	$75.41_{\pm 18.08}$	$\underline{68.98}_{\pm 15.67}$	$71.62_{\pm 21.15}$	$78.77_{\pm 13.49}$
EVA	$31.80_{\pm 17.70}$	$66.81_{\pm 9.12}$	$70.07_{\pm 6.05}$	$36.11_{\pm 14.79}$	$72.61_{\pm 8.73}$	$76.26_{\pm 6.17}$	$48.39_{\pm 19.48}$	$75.34_{\pm 8.48}$	$81.73_{\pm 5.61}$
BOSS	$42.44_{\pm 24.37}$	$57.52_{\pm 20.11}$	$\underline{79.57}_{\pm 16.62}$	$50.64_{\pm 25.12}$	$64.86_{\pm 21.86}$	$\underline{84.70}_{\pm 14.07}$	$58.11_{\pm 23.22}$	$71.36_{\pm 18.94}$	$\underline{88.95}_{\pm 9.49}$
SCLCS	$48.38_{\pm 23.00}$	$\underline{70.27}_{\pm 23.27}$	$64.18_{\pm 25.98}$	$50.59_{\pm 21.72}$	$\underline{73.86}_{\pm 15.67}$	$66.15_{\pm 21.96}$	$61.12_{\pm 16.75}$	$\underline{76.89}_{\pm 9.86}$	$68.43_{\pm 17.52}$
SCLCS$_{\text{Dense}}$	$\mathbf{57.29}_{\pm 24.07}$	$\mathbf{74.62}_{\pm 18.02}$	$\mathbf{81.87}_{\pm 9.68}$	$\mathbf{64.34}_{\pm 16.97}$	$\mathbf{77.52}_{\pm 17.18}$	$\mathbf{86.25}_{\pm 7.22}$	$\mathbf{69.84}_{\pm 14.71}$	$\mathbf{82.70}_{\pm 11.04}$	$\mathbf{89.45}_{\pm 6.81}$
SPS$_{\text{MHA}}$	$19.13_{\pm 15.36}$	$26.82_{\pm 16.34}$	$27.45_{\pm 19.31}$	$20.13_{\pm 12.27}$	$22.75_{\pm 14.33}$	$23.19_{\pm 13.22}$	$25.73_{\pm 11.92}$	$17.85_{\pm 12.03}$	$20.73_{\pm 14.90}$

MDD Diagnosis This cohort-level task requires broader structural coverage than fingerprinting. As shown in Table 3, **SCLCS**$_{\text{Dense}}$ achieves superior performance with lower variance across sampling ratios and evaluation depths, highlighting the benefit of density-aware sampling for capturing diverse patterns in group-comparison benchmarking. In this setting, the standard **SCLCS** (top-k) is less effective, suggesting that the best sampling strategy depends on the task's structural demands. Finally, simpler heuristics such as k-Means and class-imbalance–prone criteria such as Entropy (Figure 2) perform less competitively, reinforcing the need for structure-aware selection.

The behavior of the Random baseline suggests that ranking-preserving selection differs from traditional single-model core-set objectives, and that methods designed for the latter may not transfer well. The non-monotonic trend (30% < 10%) observed for **SCLCS** and **SCLCS**$_{\text{Dense}}$ is consistent with **Theorem 3**, indicating that naïve score-based top-k sampling can be brittle: it may over-represent dense clusters of typical patterns while missing rarer but important structures. This motivates **SCLCS**$_{\text{Dense}}$, which mitigates this failure mode via density-aware sampling.

Table 4: Empirical validation of **Theorem 2**. Our modified Transformer approximates a diverse set of SPIs. The model demonstrates effective fitting and generalization.

SPI Operator	**Train MSE (Start)**	**Train MSE (End)**	**Test MSE**	**SPI Operator**	**Train MSE (Start)**	**Train MSE (End)**	**Test MSE**
pec_orth	0.0605	0.0584	0.0584	plv_multitaper_mean	0.3122	0.0588	0.0585
phase_multitaper_mean	0.0295	0.0264	0.0265	cohmag_multitaper_max	0.5721	0.0055	0.0057
pli_multitaper_mean	0.0239	0.0221	0.0222	te_kernel	0.8513	0.0181	0.0182
wpli_multitaper_mean	0.0237	0.0221	0.0222	bary_euclidean_max	1.9840	0.1164	0.1153
psi_wavelet_max	6.0959	3.5522	3.5789	xme_gaussian	0.9223	0.0155	0.0158
ppc_multitaper_mean	0.2082	0.1391	0.1382	je_kernel	7.5831	0.1021	0.1025
gwtau	6.4630	1.8087	1.8124	ce_kernel	0.9243	0.0141	0.0144
icoh_multitaper_mean	0.0797	0.0206	0.0206	lcss_constraint	0.2608	0.0020	0.0020

Empirical Validation of Theorem 2 To empirically test the approximation capacity implied by **Theorem 2**, we train our modified Transformer to approximate the FC matrices produced by 16 representative SPI operators selected from the taxonomy of Cliff et al. (2023). For each target SPI, we train a separate model on fMRI time series by minimizing the mean squared error (MSE) to the SPI-generated FC matrices. Table 4 shows low final test MSE across all 16 targets, indicating that the model can closely approximate a diverse set of SPIs. Approximation fidelity varies by SPI, but the overall trend supports **Theorem 2**.

Together, **Theorem 2** and Table 4 suggest that the architecture is expressive enough to serve as a structural probe for fMRI time series. Small approximation errors (e.g., imperfect emulation of discrete statistical tests) need not invalidate our stability signal: **SPS** aims to distinguish structurally stable samples (low **SPS**) from unstable ones (high **SPS**), rather than to maximize approximation fidelity. Quantifying how approximation error propagates to ranking preservation is an important direction for future work. Full details and convergence curves are provided in **Appendix K.1**.

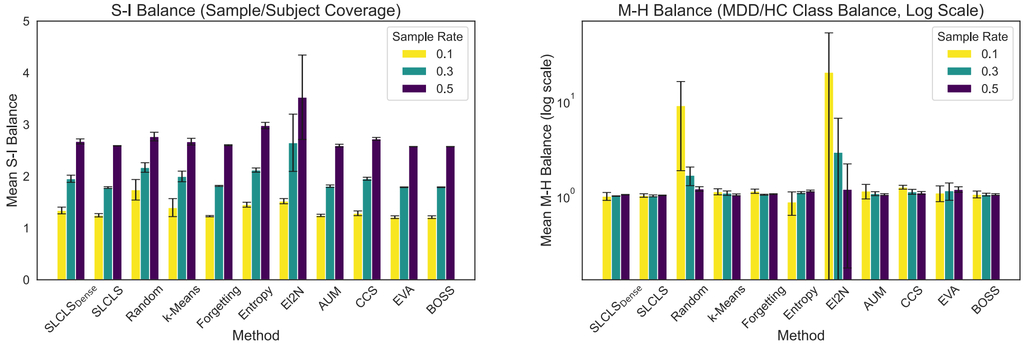

Figure 2: Sample coverage balance on subjects and MDD/HC of baselines.

Sample Coverage Balance Analysis Beyond performance, we assess the reliability and representativeness of selection by analyzing sample coverage balance (Figure 2). We introduce two metrics: *S-I Balance* (selected samples per subject, lower indicates broader subject coverage) and *M-H Balance* (MDD-to-HC ratio; deviation from 1 indicates class imbalance). As shown in Figure 2, **SCLCS** and **SCLCS$_{\textbf{Dense}}$** maintain balanced coverage with low variance on both metrics, whereas several baselines are unstable. Entropy is particularly sensitive to class-label effects, selecting almost exclusively MDD subjects at low ratios. This produces a skewed, unrepresentative core-set and can make downstream benchmarking misleading, highlighting the risk of naïve score-based selection.

5.3 Qualitative Results: Visualizing SPS Dynamics

To provide an intuitive check of the **SPS** metric, we visualize the evolution of the learned attention map $\mathbf{A}^{e}_{(\mathbf{X})}$ over the early training epochs e. While the results in **Appendix K** indicate that attention maps typically stabilize after ~ 50 epochs, Figure 3 shows that the stabilization dynamics vary across samples: a low-**SPS** sample (top row) rapidly converges to a stable structural pattern, whereas a high-**SPS** sample (bottom row) exhibits sustained fluctuations. These observations suggest that **SPS** reflects a stable, sample-specific property rather than transient optimization noise, supporting its use for identifying foundational samples for benchmarking.

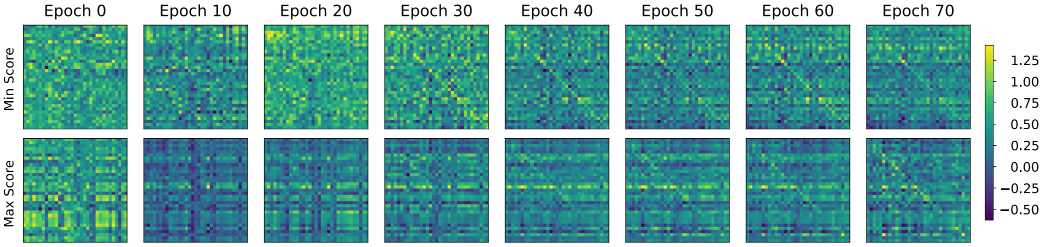

Figure 3: The evolution of the learned attention map $\mathbf{A}^{e}_{(\mathbf{X})}$ across training epochs.

6 Conclusion

In this work, we address the computational bottleneck of large-scale FC operator (SPI) benchmarking by casting core-set selection as a ranking-preservation task. Our key technical contributions are: (1) A modified Transformer architecture with a universal approximation guarantee for continuous SPI mappings under our assumptions. (2) The **SPS** metric to identify structurally stable samples. (3) The **SCLCS** framework, which outperforms 9 baselines in ranking-preservation evaluation. By accelerating FC benchmarking, **SCLCS** makes large-scale, pre-analysis SPI comparisons practical and supports more reproducible computational neuroscience.

ACKNOWLEDGEMENTS

This work is supported by Natural Science Foundation of China (No. 62402398, No. 72374173), the Fundamental Research Funds for the Central Universities (No. SWU-XDJH202303, No. SWU-KR24025), the Technological Innovation and Application Development Project of Chongqing (No. CSTB2025TIAD-KPX0027), University Innovation Research Group of Chongqing (No. CXQT21005), China Scholarship Council (CSC) program (No. 202406990056) and Chongqing Graduate Research Innovation Project (CYB240088). The experiments are supported by the High Performance Computing clusters and the Large-Scale Instrument Sharing Platform (H20 GPU Server) at Southwest University.

REPRODUCIBILITY STATEMENT

To ensure our research is fully reproducible, we have made our code, data sources, and experimental details available as follows:

- **Code:** The complete source code for our SCLCS framework and all experiments is publicly available on `https://github.com/lzhan94swu/SCLCS`.
- **Dataset and Preprocessing:** The REST-meta-MDD (Yan et al., 2019) dataset is publicly accessible at `https://rfmri.org/REST-meta-MDD`. Our detailed data preprocessing pipeline is described in **Appendix M.1**.
- **External Libraries:** Our analysis relies on the `pyspi` library (Cliff et al., 2023), which is publicly available at `https://github.com/DynamicsAndNeuralSystems/pyspi`. The specific criteria used to select SPIs for our benchmark are detailed in **Appendix H**.
- **Experimental Settings:** A summary of the experimental setup is presented in **Section 5.1**, with a comprehensive breakdown of all parameters and configurations available in **Appendix M**.
- **Theoretical Proofs:** Complete proofs for all theorems, propositions, and lemmas presented in this paper can be found in **Appendix A–G**.

ETHICS STATEMENT

Our work provides a framework for evaluating and selecting computational models used in neuroimaging analysis, which can be applied to clinical tasks such as disease diagnosis. Therefore, it is important to consider the potential societal impacts of the models ultimately chosen via our benchmarking process. A potential negative impact could arise if a core-set, while efficient, is not perfectly representative of the full dataset's diversity, leading to the selection of a model that is biased or performs suboptimally on underrepresented demographic or clinical groups. To mitigate this, we emphasize that our framework is a tool for pre-clinical scientific validation. Any model selected using our approach for real-world medical scenarios must undergo its own rigorous, independent clinical validation, and the final diagnostic decision must always remain with a qualified physician.

REFERENCES

Abhinab Acharya, Dayou Yu, Qi Yu, and Xumin Liu. Balancing feature similarity and label variability for optimal size-aware one-shot subset selection. In *Forty-first International Conference on Machine Learning*, 2024.

Elena A Allen, Eswar Damaraju, Sergey M Plis, Erik B Erhardt, Tom Eichele, and Vince D Calhoun. Tracking whole-brain connectivity dynamics in the resting state. *Cerebral cortex*, 24(3):663–676, 2014.

George D Birkhoff. Proof of the ergodic theorem. *Proceedings of the National Academy of Sciences*, 17(12):656–660, 1931.

Sebastian Bobadilla-Suarez, Christiane Ahlheim, Abhinav Mehrotra, Aristeidis Panos, and Bradley C Love. Measures of neural similarity. *Computational brain & behavior*, 3(4):369–383, 2020.

Rotem Botvinik-Nezer, Felix Holzmeister, Colin F Camerer, Anna Dreber, Juergen Huber, Magnus Johannesson, Michael Kirchler, Roni Iwanir, Jeanette A Mumford, R Alison Adcock, et al. Variability in the analysis of a single neuroimaging dataset by many teams. *Nature*, 582(7810): 84–88, 2020.

Oliver M Cliff, Annie G Bryant, Joseph T Lizier, Naotsugu Tsuchiya, and Ben D Fulcher. Unifying pairwise interactions in complex dynamics. *Nature Computational Science*, 3(10):883–893, 2023.

Cody Coleman, Christopher Yeh, Stephen Mussmann, Baharan Mirzasoleiman, Peter Bailis, Percy Liang, Jure Leskovec, and Matei Zaharia. Selection via proxy: Efficient data selection for deep learning. In *International Conference on Learning Representations*, 2020.

Open Science Collaboration. Estimating the reproducibility of psychological science. *Science*, 349 (6251):aac4716, 2015.

Nico UF Dosenbach, Binyam Nardos, Alexander L Cohen, Damien A Fair, Jonathan D Power, Jessica A Church, Steven M Nelson, Gagan S Wig, Alecia C Vogel, Christina N Lessov-Schlaggar, et al. Prediction of individual brain maturity using fmri. *Science*, 329(5997):1358–1361, 2010.

Dan Feldman. Core-sets: Updated survey. *Sampling techniques for supervised or unsupervised tasks*, pp. 23–44, 2020.

Vitaly Feldman and Chiyuan Zhang. What neural networks memorize and why: Discovering the long tail via influence estimation. *Advances in Neural Information Processing Systems*, 33:2881–2891, 2020.

Selene Gallo, Ahmed El-Gazzar, Paul Zhutovsky, Rajat M Thomas, Nooshin Javaheripour, Meng Li, Lucie Bartova, Deepti Bathula, Udo Dannlowski, Christopher Davey, et al. Functional connectivity signatures of major depressive disorder: machine learning analysis of two multicenter neuroimaging studies. *Molecular Psychiatry*, 28(7):3013–3022, 2023.

Chengcheng Guo, Bo Zhao, and Yanbing Bai. Deepcore: A comprehensive library for coreset selection in deep learning. In *International Conference on Database and Expert Systems Applications*, pp. 181–195. Springer, 2022.

John A Hartigan and Manchek A Wong. Algorithm as 136: A k-means clustering algorithm. *Journal of the royal statistical society. series c (applied statistics)*, 28(1):100–108, 1979.

Hamed Honari, Ann S Choe, and Martin A Lindquist. Evaluating phase synchronization methods in fmri: A comparison study and new approaches. *NeuroImage*, 228:117704, 2021.

Feng Hong, Yueming Lyu, Jiangchao Yao, Ya Zhang, Ivor Tsang, and Yanfeng Wang. Diversified batch selection for training acceleration. In *ICML*, 2024a.

Yuxin Hong, Xiao Zhang, Xin Zhang, and Joey Tianyi Zhou. Evolution-aware variance (eva) coreset selection for medical image classification. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 301–310, 2024b.

Jerry Yao-Chieh Hu, Hude Liu, Hong-Yu Chen, Weimin Wu, and Han Liu. Universal approximation with softmax attention. *arXiv preprint arXiv:2504.15956*, 2025.

Krishnateja Killamsetty, Sivasubramanian Durga, Ganesh Ramakrishnan, Abir De, and Rishabh Iyer. Grad-match: Gradient matching based data subset selection for efficient deep model training. In *International Conference on Machine Learning*, pp. 5464–5474. PMLR, 2021.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *ICLR*, 2015.

Ravin Kohli, Matthias Feurer, Katharina Eggensperger, Bernd Bischl, and Frank Hutter. Towards quantifying the effect of datasets for benchmarking: A look at tabular machine learning. In *ICLR Workshop*, volume 2, pp. 6, 2024.

Hojun Lee, Suyoung Kim, Junhoo Lee, Jaeyoung Yoo, and Nojun Kwak. Coreset selection for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7682–7691, 2024.

Zhixun Li, Liang Wang, Xin Sun, Yifan Luo, Yanqiao Zhu, Dingshuo Chen, Yingtao Luo, Xiangxin Zhou, Qiang Liu, Shu Wu, et al. Gslb: The graph structure learning benchmark. *Advances in Neural Information Processing Systems*, 36:30306–30318, 2023.

Shizhan Liu, Zhengkai Jiang, Yuxi Li, Jinlong Peng, Yabiao Wang, and Weiyao Lin. Density matters: improved core-set for active domain adaptive segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 13999–14007, 2024.

Zhen-Qi Liu, Andrea I Luppi, Justine Y Hansen, Ye Ella Tian, Andrew Zalesky, BT Thomas Yeo, Ben D Fulcher, and Bratislav Misic. Benchmarking methods for mapping functional connectivity in the brain. *Nature Methods*, pp. 1–10, 2025.

Yicheng Long, Hengyi Cao, Chaogan Yan, Xiao Chen, Le Li, Francisco Xavier Castellanos, Tongjian Bai, Qijing Bo, Guanmao Chen, Ningxuan Chen, et al. Altered resting-state dynamic functional brain networks in major depressive disorder: Findings from the rest-meta-mdd consortium. *NeuroImage: Clinical*, 26:102163, 2020.

Jiayu Lu, Tianyi Yan, Lan Yang, Xi Zhang, Jiaxin Li, Dandan Li, Jie Xiang, and Bin Wang. Brain fingerprinting and cognitive behavior predicting using functional connectome of high inter-subject variability. *NeuroImage*, 295:120651, 2024.

Andrea I Luppi, Helena M Gellersen, Zhen-Qi Liu, Alexander RD Peattie, Anne E Manktelow, Ram Adapa, Adrian M Owen, Lorina Naci, David K Menon, Stavros I Dimitriadis, et al. Systematic evaluation of fmri data-processing pipelines for consistent functional connectomics. *Nature Communications*, 15(1):4745, 2024.

Scott Marek, Brenden Tervo-Clemmens, Finnegan J Calabro, David F Montez, Benjamin P Kay, Alexander S Hatoum, Meghan Rose Donohue, William Foran, Ryland L Miller, Timothy J Hendrickson, et al. Reproducible brain-wide association studies require thousands of individuals. *Nature*, 603(7902):654–660, 2022.

Baharan Mirzasoleiman, Jeff Bilmes, and Jure Leskovec. Coresets for data-efficient training of machine learning models. In *International Conference on Machine Learning*, pp. 6950–6960. PMLR, 2020.

Rosaleena Mohanty, William A Sethares, Veena A Nair, and Vivek Prabhakaran. Rethinking measures of functional connectivity via feature extraction. *Scientific reports*, 10(1):1298, 2020.

Christian Otte, Stefan M Gold, Brenda W Penninx, Carmine M Pariante, Amit Etkin, Maurizio Fava, David C Mohr, and Alan F Schatzberg. Major depressive disorder. *Nature reviews Disease primers*, 2(1):1–20, 2016.

Mansheej Paul, Surya Ganguli, and Gintare Karolina Dziugaite. Deep learning on a data diet: Finding important examples early in training. *Advances in neural information processing systems*, 34: 20596–20607, 2021.

Geoff Pleiss, Tianyi Zhang, Ethan Elenberg, and Kilian Q Weinberger. Identifying mislabeled data using the area under the margin ranking. *Advances in Neural Information Processing Systems*, 33: 17044–17056, 2020.

Maria Giulia Preti, Thomas AW Bolton, and Dimitri Van De Ville. The dynamic functional connectome: State-of-the-art and perspectives. *Neuroimage*, 160:41–54, 2017.

Xiangfei Qiu, Jilin Hu, Lekui Zhou, Xingjian Wu, Junyang Du, Buang Zhang, Chenjuan Guo, Aoying Zhou, Christian S. Jensen, Zhenli Sheng, and Bin Yang. Tfb: Towards comprehensive and fair benchmarking of time series forecasting methods. In *Proc. VLDB Endow.*, pp. 2363–2377, 2024.

Lukas Roell, Stephan Wunderlich, David Roell, Florian Raabe, Elias Wagner, Zhuanghua Shi, Andrea Schmitt, Peter Falkai, Sophia Stoecklein, and Daniel Keeser. How to measure functional connectivity using resting-state fmri? a comprehensive empirical exploration of different connectivity metrics. *NeuroImage*, pp. 121195, 2025.

F Ros and S Guillaume. Core-sets: Updated survey. *Sampling Techniques for Supervised or Unsupervised Tasks*, pp. 23, 2019.

Philip Sedgwick. Spearman's rank correlation coefficient. *Bmj*, 349, 2014.

Ozan Sener and Silvio Savarese. Active learning for convolutional neural networks: A core-set approach. In *International Conference on Learning Representations*, 2018.

Stephen M Smith, Karla L Miller, Gholamreza Salimi-Khorshidi, Matthew Webster, Christian F Beckmann, Thomas E Nichols, Joseph D Ramsey, and Mark W Woolrich. Network modelling methods for fmri. *Neuroimage*, 54(2):875–891, 2011.

Stephen M Smith, Diego Vidaurre, Christian F Beckmann, Matthew F Glasser, Mark Jenkinson, Karla L Miller, Thomas E Nichols, Emma C Robinson, Gholamreza Salimi-Khorshidi, Mark W Woolrich, et al. Functional connectomics from resting-state fmri. *Trends in cognitive sciences*, 17 (12):666–682, 2013.

Mariya Toneva, Alessandro Sordoni, Remi Tachet des Combes, Adam Trischler, Yoshua Bengio, and Geoffrey J Gordon. An empirical study of example forgetting during deep neural network learning. *arXiv preprint arXiv:1812.05159*, 2018.

Mariya Toneva, Alessandro Sordoni, Remi Tachet des Combes, Adam Trischler, Yoshua Bengio, and Geoffrey J. Gordon. An empirical study of example forgetting during deep neural network learning. In *International Conference on Learning Representations*, 2019.

Dimitri Van De Ville, Younes Farouj, Maria Giulia Preti, Raphaël Liégeois, and Enrico Amico. When makes you unique: Temporality of the human brain fingerprint. *Science advances*, 7(42):eabj0751, 2021.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Yining Wang, Liwei Wang, Yuanzhi Li, Di He, and Tie-Yan Liu. A theoretical analysis of ndcg type ranking measures. In *Conference on learning theory*, pp. 25–54. PMLR, 2013.

Stanisław Węglarczyk. Kernel density estimation and its application. In *ITM web of conferences*, volume 23, pp. 00037. EDP Sciences, 2018.

Kai Wei, Rishabh Iyer, and Jeff Bilmes. Submodularity in data subset selection and active learning. In *International conference on machine learning*, pp. 1954–1963. PMLR, 2015.

Xiaobo Xia, Jiale Liu, Jun Yu, Xu Shen, Bo Han, and Tongliang Liu. Moderate coreset: A universal method of data selection for real-world data-efficient deep learning. In *The Eleventh International Conference on Learning Representations*, 2022.

Chao-Gan Yan, Xiao Chen, Le Li, Francisco Xavier Castellanos, Tong-Jian Bai, Qi-Jing Bo, Jun Cao, Guan-Mao Chen, Ning-Xuan Chen, Wei Chen, et al. Reduced default mode network functional connectivity in patients with recurrent major depressive disorder. *Proceedings of the National Academy of Sciences*, 116(18):9078–9083, 2019.

Jiahao Ying, Yixin Cao, Yushi Bai, Qianru Sun, Bo Wang, Wei Tang, Zhaojun Ding, Yizhe Yang, Xuanjing Huang, and Shuicheng Yan. Automating dataset updates towards reliable and timely evaluation of large language models. *Advances in Neural Information Processing Systems*, 37: 17106–17132, 2024.

Haizhong Zheng, Rui Liu, Fan Lai, and Atul Prakash. Coverage-centric coreset selection for high pruning rates. *arXiv preprint arXiv:2210.15809*, 2022.

Haizhong Zheng, Rui Liu, Fan Lai, and Atul Prakash. Coverage-centric coreset selection for high pruning rates. In *The Eleventh International Conference on Learning Representations*, 2023.

S Kevin Zhou, Hayit Greenspan, Christos Davatzikos, James S Duncan, Bram Van Ginneken, Anant Madabhushi, Jerry L Prince, Daniel Rueckert, and Ronald M Summers. A review of deep learning in medical imaging: Imaging traits, technology trends, case studies with progress highlights, and future promises. *Proceedings of the IEEE*, 109(5):820–838, 2021.

Zhiyao Zhou, Sheng Zhou, Bochao Mao, Xuanyi Zhou, Jiawei Chen, Qiaoyu Tan, Daochen Zha, Yan Feng, Chun Chen, and Can Wang. Opengsl: A comprehensive benchmark for graph structure learning. *Advances in Neural Information Processing Systems*, 36:17904–17928, 2023.

Yongcheng Zong, Qiankun Zuo, Michael Kwok-Po Ng, Baiying Lei, and Shuqiang Wang. A new brain network construction paradigm for brain disorder via diffusion-based graph contrastive learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

A Proof of Theorem 1

Theorem (Interference of Averaged Attention, full version)**.** *Let $\{\mathbf{A}_h\}_{h=1}^H$ be a collection of row stochastic*[2] *attention matrices with $\mathbf{A}_h \in \mathbb{R}^{N\times N}$. For every row index $i \in \{1,\dots,N\}$ assume there exist sets $S_h^{(i)} \subseteq \{1,\dots,N\}$ that are pairwise disjoint, meaning $S_h^{(i)} \cap S_{h'}^{(i)} = \varnothing$ for all $h \neq h'$, and such that*

$$\mathbf{A}_h(i,j) = 0 \quad \textit{for all } j \notin S_h^{(i)}.$$

Define the uniform average $\bar{\mathbf{A}} := \frac{1}{H}\sum_{h=1}^H \mathbf{A}_h$. Then for every row i:

(a) ***Support expansion.*** $\operatorname{supp}(\bar{\mathbf{a}}^{(i)}) = \bigcup_{h=1}^H S_h^{(i)}$*. If $H \geq 2$, then for every h,* $\operatorname{supp}(\bar{\mathbf{a}}^{(i)}) \not\subseteq S_h^{(i)}$*.*

(b) ***Entropy inflation.*** *With $\mathcal{H}(\mathbf{p}) := -\sum_{j=1}^N p_j \log p_j$,*

$$\mathcal{H}\big(\bar{\mathbf{a}}^{(i)}\big) \;\geq\; \frac{1}{H}\sum_{h=1}^H \mathcal{H}\big(\mathbf{a}_h^{(i)}\big) \;\geq\; \min_{1\leq h\leq H} \mathcal{H}\big(\mathbf{a}_h^{(i)}\big),$$

and strict inequality holds whenever the vectors $\{\mathbf{a}_h^{(i)}\}_{h=1}^H$ are not all identical. In particular, under the disjoint mask assumption, strict inequality holds for every i whenever $H \geq 2$.

Proof. Fix a row index i and write $\mathbf{p}_h := \mathbf{a}_h^{(i)} \in \Delta^{N-1}$. By definition,

$$\bar{\mathbf{p}} := \bar{\mathbf{a}}^{(i)} = \frac{1}{H}\sum_{h=1}^H \mathbf{p}_h, \qquad \bar{\mathbf{p}}_j = \frac{1}{H}\sum_{h=1}^H (\mathbf{p}_h)_j.$$

(a) Support expansion. All entries are nonnegative, so $\bar{\mathbf{p}}_j > 0$ if and only if there exists a head h with $(\mathbf{p}_h)_j > 0$. Under the masking assumption, $(\mathbf{p}_h)_j > 0$ implies $j \in S_h^{(i)}$. Conversely, since each $\mathbf{p}_h$ is a probability vector supported on $S_h^{(i)}$, every $j \in S_h^{(i)}$ with nonzero mass in head h contributes a positive term to $\bar{\mathbf{p}}_j$. Therefore,

$$\operatorname{supp}(\bar{\mathbf{p}}) = \bigcup_{h=1}^H \operatorname{supp}(\mathbf{p}_h) \subseteq \bigcup_{h=1}^H S_h^{(i)}.$$

Moreover, because $\mathbf{p}_h$ is row stochastic and supported inside $S_h^{(i)}$, we have $\operatorname{supp}(\mathbf{p}_h) \neq \varnothing$ and $\operatorname{supp}(\mathbf{p}_h) \subseteq S_h^{(i)}$, so $\operatorname{supp}(\bar{\mathbf{p}})$ equals the union of the head supports and is contained in the union of the masks. If, as in the theorem statement, the intended structural supports are exactly the mask sets

[2] Each row is a probability distribution produced by a softmax: $\mathbf{A}_h(i,j) \geq 0$ and $\sum_{j=1}^N \mathbf{A}_h(i,j) = 1$.

(that is, the mask defines which indices can receive positive mass), then $\operatorname{supp}(\mathbf{p}_h) = S_h^{(i)}$ and hence $\operatorname{supp}(\bar{\mathbf{p}}) = \bigcup_{h=1}^{H} S_h^{(i)}$. When $H \geq 2$ and the sets $\{S_h^{(i)}\}_{h=1}^{H}$ are pairwise disjoint, the union strictly contains each $S_h^{(i)}$, so $\operatorname{supp}(\bar{\mathbf{p}}) \not\subseteq S_h^{(i)}$ for every h.

(b) Entropy inflation. The Shannon entropy $\mathcal{H}$ is strictly concave on the probability simplex. By Jensen's inequality,

$$\mathcal{H}(\bar{\mathbf{p}}) \;\geq\; \frac{1}{H}\sum_{h=1}^{H} \mathcal{H}(\mathbf{p}_h) \;\geq\; \min_{1\leq h\leq H} \mathcal{H}(\mathbf{p}_h),$$

and the first inequality is strict unless $\mathbf{p}_1 = \cdots = \mathbf{p}_H$. Under the disjoint mask assumption with $H \geq 2$, the vectors cannot all be identical: if $\mathbf{p}_h = \mathbf{p}_{h'}$ for some $h \neq h'$, then their supports coincide and are nonempty, so $S_h^{(i)} \cap S_{h'}^{(i)} \neq \varnothing$, contradicting disjointness. Hence $\mathbf{p}_1, \ldots, \mathbf{p}_H$ are not all identical, so $\mathcal{H}(\bar{\mathbf{p}}) > \frac{1}{H}\sum_{h=1}^{H} \mathcal{H}(\mathbf{p}_h)$ and therefore $\mathcal{H}(\bar{\mathbf{p}}) > \min_h \mathcal{H}(\mathbf{p}_h)$.

Together, (a) and (b) show that uniform averaging introduces additional nonzero entries and increases entropy, which blurs head specific structural patterns. □

B PROOF OF THEOREM 2

Proof. We use an existing attention-only universal approximation result as the main engine and then specialize it to row-stochastic matrix-valued targets.

Step 1 (Reduce to sequence-to-sequence approximation). View $\mathbf{X} \in \mathbb{R}^{N\times T}$ as a length-N sequence with token dimension T. Likewise, view $S(\mathbf{X}) \in \mathbb{R}^{N\times N}$ as a length-N sequence of row vectors in $\mathbb{R}^N$. Since $\mathcal{X}$ is compact and S is continuous, this is a continuous sequence-to-sequence map on a compact domain.

Step 2 (Attention-only universality on compact sets). Recent results show that (softmax-)attention-only architectures with linear projections are universal approximators for continuous sequence-to-sequence maps on compact domains (Hu et al., 2025). In particular, attention modules can simulate piecewise-linear bases and hence achieve uniform approximation without requiring feed-forward sublayers. We invoke such a theorem.

Step 3 (Constrain the output to be row-stochastic). Our target operator satisfies $S(\mathbf{X}) \in \Delta^{N-1\times N}$ for all $\mathbf{X}$. The model family in equation 2 is also row-stochastic by construction: each head output is row-stochastic (row-wise softmax), and the convex mixture with $\boldsymbol{\alpha} \in \Delta^{H-1}$ preserves row-stochasticity. Therefore the approximating attention-only construction can be chosen to lie entirely inside $\Delta^{N-1\times N}$.

Step 4 (Uniform approximation in Frobenius norm). The cited attention-only universality provides uniform approximation in a sup norm over the compact domain. Since $\|\cdot\|_F \leq \sqrt{N}\|\cdot\|_{\infty,\infty}$, the same parameter choice yields

$$\sup_{\mathbf{X}\in\mathcal{X}} \big\|\mathbf{A}_\theta(\mathbf{X}) - S(\mathbf{X})\big\|_F < \varepsilon$$

after tightening constants.

This completes the proof. □

C PROOF OF PROPOSITION 1

Proof. Because Z_e and Z_{e-1} are i.i.d.,

$$\begin{aligned}\mathbb{E}[\Delta_e] &= \sum_{k=1}^{K}\sum_{l=1}^{K} \Pr[Z_e = S^{(k)},\, Z_{e-1} = S^{(l)}]\, \|S^{(k)} - S^{(l)}\|_F^2 \\ &= \sum_{k=1}^{K}\sum_{l=1}^{K} \lambda_k \lambda_l D_{kl},\end{aligned}$$

which gives the first equality in equation 7. Since $D_{kk} = 0$ and $D_{kl} = D_{lk}$,

$$\sum_{k=1}^{K}\sum_{l=1}^{K} \lambda_k \lambda_l D_{kl} = 2 \sum_{k<l} \lambda_k \lambda_l D_{kl},$$

proving the second equality in equation 7.

For the bounds equation 8, note that for all $k < l$, $D_{\min} \le D_{kl} \le D_{\max}$, hence

$$2D_{\min} \sum_{k<l} \lambda_k \lambda_l \;\le\; 2 \sum_{k<l} \lambda_k \lambda_l D_{kl} \;\le\; 2D_{\max} \sum_{k<l} \lambda_k \lambda_l.$$

Finally,

$$2 \sum_{k<l} \lambda_k \lambda_l = \Big(\sum_{k=1}^{K} \lambda_k\Big)^2 - \sum_{k=1}^{K} \lambda_k^2 = 1 - \sum_{k=1}^{K} \lambda_k^2,$$

which is the Gini impurity of $\{\lambda_k\}$. This yields equation 8.

If $D_{kl} \equiv D$ for all $k \neq l$, then equation 7 becomes

$$\mathbb{E}[\Delta_e] = 2D \sum_{k<l} \lambda_k \lambda_l = D\Big(1 - \sum_{k=1}^{K} \lambda_k^2\Big),$$

establishing equation 9. □

D Lemma 1 and Proof

Lemma 1 (Consistency of SPS). *Let $\{\mathbf{A}^{(e)}_{(\mathbf{X})}\}_{e \ge 0}$ be the sequence of attention-based structure matrices for a fixed sample $\mathbf{X}$ generated by a stochastic optimization algorithm. Assume the sequence of differences*

$$\Delta_e(\mathbf{X}) \;=\; \big\|\mathbf{A}^{(e)}_{(\mathbf{X})} - \mathbf{A}^{(e-1)}_{(\mathbf{X})}\big\|_F^2 \tag{18}$$

forms a stationary and ergodic *process with finite mean $\sigma^2(\mathbf{X}) = \mathbb{E}[\Delta_e(\mathbf{X})]$. Then the empirical SPS estimator*

$$\widehat{\mathrm{SPS}}_L(\mathbf{X}) \;=\; \frac{1}{L} \sum_{e=1}^{L} \Delta_e(\mathbf{X}) \tag{19}$$

converges almost surely to $\sigma^2(\mathbf{X})$ as $L \to \infty$: $\widehat{\mathrm{SPS}}_L(\mathbf{X}) \;\longrightarrow\; \sigma^2(\mathbf{X}) \quad$ a.s.

Proof. Because $\{\Delta_e(\mathbf{X})\}$ is assumed stationary and ergodic with finite first moment, Birkhoff's pointwise ergodic theorem (Birkhoff, 1931) applies:

$$\frac{1}{L} \sum_{e=1}^{L} \Delta_e(\mathbf{X}) \;\xrightarrow{\text{a.s.}}\; \mathbb{E}[\Delta_e(\mathbf{X})] \;=\; \sigma^2(\mathbf{X}). \tag{20}$$

But the left–hand side is precisely $\widehat{\mathrm{SPS}}_L(\mathbf{X})$. Hence the estimator is strongly consistent. □

E Proof of Theorem 3

Proof. Write the scores as $\{s_i\}_{i=1}^N$ and let $s_{(k)}$ be the k-th order statistic. Equivalently, $S_k = \{\mathbf{x}_i : s_i \le s_{(k)}\}$ up to tie-breaking on a null event (because the score distributions are continuous).

Step 1: Identify the population selection threshold. Let τ satisfy equation 11. Intuitively, τ is the population score threshold whose expected accepted fraction is ρ:

$$\mathbb{E}\Big[\frac{1}{N} \sum_{i=1}^{N} \mathbb{1}\{s_i \le \tau\}\Big] = \pi_p F_p(\tau) + \pi_q F_q(\tau) = \rho.$$

Step 2: The empirical threshold concentrates. Define the empirical accepted fraction at threshold t:

$$G_N(t) := \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}\{s_i \le t\}.$$

Because scores are independent across samples and each indicator is bounded, a standard concentration argument (e.g., Hoeffding) yields that $G_N(t)$ concentrates uniformly on compact intervals around its mean $G(t) := \pi_p F_p(t) + \pi_q F_q(t)$. Since F_p, F_q are continuous and G is strictly increasing at τ (under mild regularity), it follows that the empirical quantile $s_{(k)}$ converges in probability to τ:

$$s_{(k)} \xrightarrow[N\to\infty]{\Pr} \tau. \tag{21}$$

Step 3: Selected cluster counts converge to their expectations. Conditional on $s_{(k)}$, the number of selected points from cluster C_p is

$$|S_k \cap C_p| = \sum_{\mathbf{x}\in C_p} \mathbb{1}\{s(\mathbf{x}) \le s_{(k)}\}.$$

Given $s_{(k)}$, the indicators in the sum are i.i.d. Bernoulli with parameter $F_p(s_{(k)})$, so by the law of large numbers (and Slutsky using equation 21),

$$\frac{|S_k \cap C_p|}{n_p} \xrightarrow{\Pr} F_p(\tau). \tag{22}$$

Similarly,

$$\frac{|S_k \cap C_q|}{n_q} \xrightarrow{\Pr} F_q(\tau). \tag{23}$$

Step 4: Convert to selected proportions and compute the bias. Divide equation 22 by $k/N \to \rho$:

$$\widehat{\pi}_p(S_k) = \frac{|S_k \cap C_p|}{k} = \frac{|S_k \cap C_p|/N}{k/N} = \frac{(n_p/N)\cdot(|S_k \cap C_p|/n_p)}{k/N} \xrightarrow{\Pr} \frac{\pi_p F_p(\tau)}{\rho}.$$

Using equation 11, we compute

$$\frac{\pi_p F_p(\tau)}{\rho} - \pi_p = \pi_p\Big(\frac{F_p(\tau)-\rho}{\rho}\Big) = \pi_p\Big(\frac{F_p(\tau) - \pi_p F_p(\tau) - \pi_q F_q(\tau)}{\rho}\Big) = \frac{\pi_p\pi_q}{\rho}\big(F_p(\tau) - F_q(\tau)\big).$$

Under equation 12, this equals $\delta := \frac{\pi_p\pi_q}{\rho}\gamma > 0$, establishing the first limit in equation 13. The statement for $\widehat{\pi}_q(S_k)$ follows since $\widehat{\pi}_q(S_k) = 1 - \widehat{\pi}_p(S_k)$, and then $\Delta_k = 2|\widehat{\pi}_p(S_k) - \pi_p| \xrightarrow{\Pr} 2\delta > 0$. □

F Theorem 4 and Proof

Theorem 4 (ε-coverage of density-reweighted sampling). *Fix $0 < \delta < 1$ and $\varepsilon > 0$. Let $\tilde{\mathcal{X}}_c$ be the candidate pool with $n = |\tilde{\mathcal{X}}_c|$, and let $S \subset \tilde{\mathcal{X}}_c$ be the subset of size m returned by the proposed sampling procedure. Let N_ε be the ε-covering number of $\mathcal{X}_c$ under $d(\mathbf{A}, \mathbf{A}') = \|\mathbf{A} - \mathbf{A}'\|_F$. Under* **Assumption 1***, if*

$$m \ \ge\ \frac{n(\rho_{\max} + \tau)}{\rho_{\min} + \tau}\Big(\log N_\varepsilon + \log(1/\delta)\Big), \tag{24}$$

then S is an ε-cover of $\mathcal{X}_c$ with probability at least $1 - \delta$.

Define the structure-representation space $(\mathcal{M}, d)$ with metric $d(\mathbf{A}, \mathbf{A}') = \|\mathbf{A} - \mathbf{A}'\|_F$. For $\varepsilon > 0$ let N_ε be the covering number of $\mathcal{X}_c \subset \mathcal{M}$. That is, the smallest number of closed d-balls of radius ε needed to cover all subjects in $\mathcal{X}_c$.

Assumption 1. *After the β-filter step, define the KDE-induced density $\rho(\mathbf{X}) := \hat{p}_{\mathrm{SPS}}\big(\mathrm{SPS}(\mathbf{X})\big)$, where $\hat{p}_{\mathrm{SPS}}$ is a Gaussian KDE fit on $\{\mathrm{SPS}(\mathbf{X}) : \mathbf{X} \in \tilde{\mathcal{X}}_c\}$. Assume the estimator is bounded on the candidate pool: $0 < \rho_{\min} \le \rho(\mathbf{X}) \le \rho_{\max} < \infty$ for every $\mathbf{X} \in \tilde{\mathcal{X}}_c$.*

Assumption 1 is mild because Gaussian KDE with finite bandwidth produces a bounded, strictly positive estimate on any finite sample.

Proof. Define the (normalized) sampling weight for $\mathbf{X} \in \tilde{\mathcal{X}}_c$ as

$$w(\mathbf{X}) = \frac{\frac{1}{\rho(\mathbf{X})+\tau}}{\sum_{\mathbf{Z}\in\tilde{\mathcal{X}}_c} \frac{1}{\rho(\mathbf{Z})+\tau}}. \tag{25}$$

By **Assumption 1**, for all $\mathbf{X}$, $\frac{1}{\rho(\mathbf{X})+\tau} \geq \frac{1}{\rho_{\max}+\tau}$ and $\sum_{\mathbf{Z}} \frac{1}{\rho(\mathbf{Z})+\tau} \leq \frac{n}{\rho_{\min}+\tau}$. Hence every point has weight bounded below by

$$w(\mathbf{X}) \;\geq\; \frac{1/(\rho_{\max}+\tau)}{n/(\rho_{\min}+\tau)} = \frac{\rho_{\min}+\tau}{n(\rho_{\max}+\tau)} \;=:\; w_{\min}. \tag{26}$$

Let $\{B_1, \ldots, B_{N_\varepsilon}\}$ be a collection of closed d-balls of radius ε covering $\mathcal{X}_c$. Since $\mathcal{X}_c \subseteq \tilde{\mathcal{X}}_c$, each B_j contains at least one candidate point. Therefore its total sampling mass satisfies $W_j := \sum_{\mathbf{X}\in B_j\cap\tilde{\mathcal{X}}_c} w(\mathbf{X}) \geq w_{\min}$.

Consider sequential sampling without replacement where at each draw we sample from the remaining points proportionally to their weights (renormalized). If we have not yet sampled from B_j, then removing points outside B_j can only increase the renormalized mass of B_j; thus, at every draw the conditional probability of selecting a point outside B_j is at most $1 - W_j \leq 1 - w_{\min}$. Therefore

$$\Pr[B_j \cap S = \varnothing] \;\leq\; (1 - w_{\min})^m \;\leq\; \exp(-w_{\min} m). \tag{27}$$

Choose m so that $\exp(-w_{\min} m) \leq \delta/N_\varepsilon$, i.e. $m \geq \frac{1}{w_{\min}}(\log N_\varepsilon + \log(1/\delta))$. A union bound over the N_ε balls yields $\Pr[\exists j : B_j \cap S = \varnothing] \leq \delta$. Hence with probability at least $1-\delta$, every ball contains at least one sampled point, so S is an ε-cover of $\mathcal{X}_c$. □

G Theorem 5 and Proof

Theorem 5 (Expectation discrepancy under ε-coverage)**.** *Let $(\mathcal{M}, d)$ be the structure-representation space with $d(\mathbf{A}, \mathbf{A}') = \|\mathbf{A}-\mathbf{A}'\|_F$ and representation map $\mathbf{X} \mapsto \mathbf{A}^{(\mathbf{X})} \in \mathcal{M}$. Let P be a probability distribution supported on $\mathcal{X}$. Assume $\tilde{\mathcal{X}} \subset \mathcal{X}$ is an ε-cover of $\mathcal{X}$ in $\mathcal{M}$: for every $\mathbf{X} \in \mathcal{X}$ there exists $\tilde{\mathbf{X}} \in \tilde{\mathcal{X}}$ with $d(\mathbf{A}^{(\mathbf{X})}, \mathbf{A}^{(\tilde{\mathbf{X}})}) \leq \varepsilon$.*

Let $\pi : \mathcal{X} \to \tilde{\mathcal{X}}$ be any (measurable) selection satisfying $d(\mathbf{A}^{(\mathbf{X})}, \mathbf{A}^{(\pi(\mathbf{X}))}) \leq \varepsilon$ for all $\mathbf{X}$, and define the push-forward measure $P_\pi := P \circ \pi^{-1}$ on $\tilde{\mathcal{X}}$. Then for any function $f : \mathcal{X} \to \mathbb{R}$ that is L-Lipschitz w.r.t. d,

$$\left|\mathbb{E}_{\mathbf{X}\sim P}\big[f(\mathbf{X})\big] - \mathbb{E}_{\mathbf{X}\sim P_\pi}\big[f(\mathbf{X})\big]\right| \leq L\,\varepsilon. \tag{28}$$

Proof. Because f is L-Lipschitz w.r.t. d, for every $\mathbf{X} \in \mathcal{X}$,

$$\big|f(\mathbf{X}) - f(\pi(\mathbf{X}))\big| \leq L\, d\big(\mathbf{A}^{(\mathbf{X})}, \mathbf{A}^{(\pi(\mathbf{X}))}\big) \leq L\,\varepsilon.$$

Taking expectation under P and using Jensen/triangle inequality,

$$\left|\mathbb{E}_{\mathbf{X}\sim P}\big[f(\mathbf{X})\big] - \mathbb{E}_{\mathbf{X}\sim P}\big[f(\pi(\mathbf{X}))\big]\right| \leq \mathbb{E}_{\mathbf{X}\sim P}\big[|f(\mathbf{X}) - f(\pi(\mathbf{X}))|\big] \leq L\,\varepsilon.$$

Finally, by definition of the push-forward measure P_π,

$$\mathbb{E}_{\mathbf{X}\sim P}\big[f(\pi(\mathbf{X}))\big] = \mathbb{E}_{\mathbf{X}\sim P_\pi}\big[f(\mathbf{X})\big].$$

Combining completes the proof. □

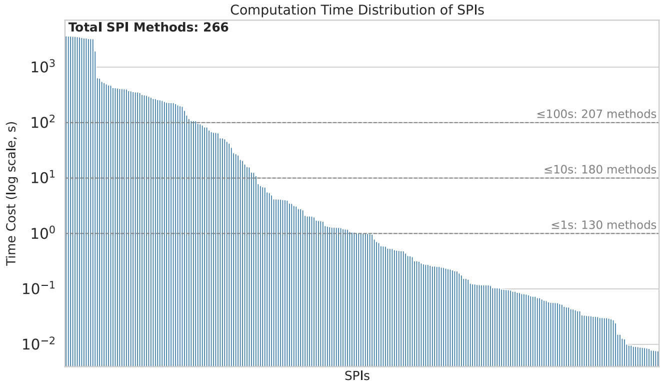

Figure A1: Time consumption of different SPIs on a single sample.

Table A1: SPIs included in `pyspi` (**bolded** ones are used in this paper).

cov_EllipticEnvelope	**cov_GraphicalLasso**	**cov_LedoitWolf**	**cov_MinCovDet**
cov_OAS	**cov_ShrunkCovariance**	**cov-sq_EmpiricalCovariance**	**cov-sq_EllipticEnvelope**
cov-sq_GraphicalLasso	**cov-sq_LedoitWolf**	**cov-sq_MinCovDet**	**cov-sq_OAS**
cov_PearsonCorrelation	**cov-sq_ShrunkCovariance**	**prec_EmpiricalCovariance**	**prec_EllipticEnvelope**
prec_GraphicalLasso	**prec_LedoitWolf**	**prec_MinCovDet**	**prec_OAS**
prec_ShrunkCovariance	**prec-sq_EmpiricalCovariance**	**prec-sq_EllipticEnvelope**	**prec-sq_GraphicalLasso**
prec-sq_LedoitWolf	**prec-sq_MinCovDet**	**prec-sq_OAS**	**prec-sq_ShrunkCovariance**
kendalltau-sq	**kendalltau**	**xcorr_max_sig-True**	**xcorr-sq_max_sig-True**
xcorr_mean_sig-True	**xcorr-sq_mean_sig-True**	**xcorr_mean_sig-False**	**xcorr-sq_mean_sig-False**
pdist_cityblock	**pdist_cosine**	**pdist_chebyshev**	**pdist_canberra**
pdist_braycurtis	**lcss_constraint-sakoe-chiba**	**bary_euclidean_mean**	**bary_euclidean_max**
bary-sq_euclidean_mean	**bary-sq_euclidean_max**	**gwtau**	**je_kernel_W-0.5**
ce_gaussian	**ce_kernel_W-0.5**	**xme_gaussian_k1**	**xme_gaussian_k10**
mi_gaussian	**tlmi_gaussian**	**te_kernel_W-0.25_k-1**	**gc_gaussian_k-max-10_tau-max-2**
gc_gaussian_k-1_kt-1_l-1_lt-1	**te_symbolic_k-1_kt-1_l-1_lt-1**	**te_symbolic_k-10_kt-1_l-1_lt-1**	**phase_multitaper_mean_fs-1_fmin-0_fmax-0-5**
phase_multitaper_mean_fs-1_fmin-0_fmax-0-25	**phase_multitaper_mean_fs-1_fmin-0-25_fmax-0-5**	**phase_multitaper_max_fs-1_fmin-0_fmax-0-5**	**phase_multitaper_max_fs-1_fmin-0_fmax-0-25**
phase_multitaper_max_fs-1_fmin-0-25_fmax-0-5	**cohmag_multitaper_mean_fs-1_fmin-0_fmax-0-5**	**cohmag_multitaper_mean_fs-1_fmin-0_fmax-0-25**	**cohmag_multitaper_mean_fs-1_fmin-0-25_fmax-0-5**
cohmag_multitaper_max_fs-1_fmin-0_fmax-0-5	**cohmag_multitaper_max_fs-1_fmin-0_fmax-0-25**	**cohmag_multitaper_max_fs-1_fmin-0-25_fmax-0-5**	**icoh_multitaper_mean_fs-1_fmin-0_fmax-0-5**
icoh_multitaper_mean_fs-1_fmin-0_fmax-0-25	**icoh_multitaper_mean_fs-1_fmin-0-25_fmax-0-5**	**icoh_multitaper_max_fs-1_fmin-0_fmax-0-5**	**icoh_multitaper_max_fs-1_fmin-0_fmax-0-25**
icoh_multitaper_max_fs-1_fmin-0-25_fmax-0-5	**psi_multitaper_mean_fs-1_fmin-0_fmax-0-5**	**psi_multitaper_mean_fs-1_fmin-0_fmax-0-25**	**psi_multitaper_mean_fs-1_fmin-0-25_fmax-0-5**
plv_multitaper_mean_fs-1_fmin-0_fmax-0-5	**plv_multitaper_mean_fs-1_fmin-0_fmax-0-25**	**plv_multitaper_mean_fs-1_fmin-0-25_fmax-0-5**	**plv_multitaper_max_fs-1_fmin-0_fmax-0-5**
plv_multitaper_max_fs-1_fmin-0_fmax-0-25	**plv_multitaper_max_fs-1_fmin-0-25_fmax-0-5**	**pli_multitaper_mean_fs-1_fmin-0_fmax-0-5**	**pli_multitaper_mean_fs-1_fmin-0_fmax-0-25**
pli_multitaper_mean_fs-1_fmin-0-25_fmax-0-5	**pli_multitaper_max_fs-1_fmin-0_fmax-0-5**	**pli_multitaper_max_fs-1_fmin-0_fmax-0-25**	**pli_multitaper_max_fs-1_fmin-0-25_fmax-0-5**
wpli_multitaper_mean_fs-1_fmin-0_fmax-0-5	**wpli_multitaper_mean_fs-1_fmin-0_fmax-0-25**	**wpli_multitaper_mean_fs-1_fmin-0-25_fmax-0-5**	**wpli_multitaper_max_fs-1_fmin-0_fmax-0-5**
wpli_multitaper_max_fs-1_fmin-0_fmax-0-25	**wpli_multitaper_max_fs-1_fmin-0-25_fmax-0-5**	**dspli_multitaper_mean_fs-1_fmin-0_fmax-0-5**	**dspli_multitaper_mean_fs-1_fmin-0_fmax-0-25**
dspli_multitaper_mean_fs-1_fmin-0-25_fmax-0-5	**dspli_multitaper_max_fs-1_fmin-0_fmax-0-5**	**dspli_multitaper_max_fs-1_fmin-0_fmax-0-25**	**dspli_multitaper_max_fs-1_fmin-0-25_fmax-0-5**
dswpli_multitaper_mean_fs-1_fmin-0_fmax-0-5	**dswpli_multitaper_mean_fs-1_fmin-0_fmax-0-25**	**dswpli_multitaper_mean_fs-1_fmin-0-25_fmax-0-5**	**dswpli_multitaper_max_fs-1_fmin-0_fmax-0-5**
dswpli_multitaper_max_fs-1_fmin-0_fmax-0-25	**dswpli_multitaper_max_fs-1_fmin-0-25_fmax-0-5**	**ppc_multitaper_mean_fs-1_fmin-0_fmax-0-5**	**ppc_multitaper_mean_fs-1_fmin-0_fmax-0-25**
ppc_multitaper_mean_fs-1_fmin-0-25_fmax-0-5	**ppc_multitaper_max_fs-1_fmin-0_fmax-0-5**	**ppc_multitaper_max_fs-1_fmin-0_fmax-0-25**	**ppc_multitaper_max_fs-1_fmin-0-25_fmax-0-5**
gd_multitaper_delay_fs-1_fmin-0_fmax-0-5	**gd_multitaper_delay_fs-1_fmin-0-25_fmax-0-5**	**sgc_parametric_mean_fs-1_fmin-0_fmax-0-25_order-1**	**sgc_parametric_max_fs-1_fmin-1e-05_fmax-0-5_order-1**
psi_wavelet_mean_fs-1_fmin-0_fmax-0-25_mean	**psi_wavelet_mean_fs-1_fmin-0-25_fmax-0-5_mean**	**psi_wavelet_max_fs-1_fmin-0_fmax-0-5_max**	**psi_wavelet_max_fs-1_fmin-0_fmax-0-25_max**
psi_wavelet_max_fs-1_fmin-0-25_fmax-0-5_max	**pec**	**pec_orth**	**pec_log**
pec_orth_log	**pec_orth_abs**	**pec_orth_log_abs**	cov_GraphicalLassoCV
cov-sq_GraphicalLassoCV	prec_GraphicalLassoCV	prec-sq_GraphicalLassoCV	spearmanr-sq
spearmanr	pdist_euclidean	dcorr	dcorr_biased
mgc	hsic	hsic_biased	hhg
mgcx_maxlag-1	mgcx_maxlag-10	dcorrx_maxlag-1	dcorrx_maxlag-10
dtw	dtw_constraint-itakura	dtw_constraint-sakoe-chiba	softdtw
softdtw_constraint-itakura	softdtw_constraint-sakoe-chiba	lcss	lcss_constraint-itakura
bary_dtw_mean	bary_dtw_max	bary_sgddtw_mean	bary_sgddtw_max
bary_softdtw_mean	bary_softdtw_max	bary-sq_dtw_mean	bary-sq_dtw_max
bary-sq_sgddtw_mean	bary-sq_sgddtw_max	bary-sq_softdtw_mean	bary-sq_softdtw_max
anm	cds	reci	igci
ccm_E-None_mean	ccm_E-None_max	ccm_E-None_diff	ccm_E-1_mean
ccm_E-1_max	ccm_E-1_diff	ccm_E-10_mean	ccm_E-10_max
ccm_E-10_diff	je_gaussian	je_kozachenko	ce_kozachenko
cce_gaussian	cce_kozachenko	cce_kernel_W-0.5	xme_kozachenko_k1
xme_kernel_W-0.5_k1	xme_kozachenko_k10	xme_kernel_W-0.5_k10	di_gaussian
di_kozachenko	di_kernel_W-0.5	si_gaussian_k-1	si_kozachenko_k-1
si_kernel_W-0.5_k-1	mi_kraskov_NN-4	mi_kraskov_NN-4_DCE	mi_kernel_W-0.25
tlmi_kraskov_NN-4	tlmi_kraskov_NN-4_DCE	tlmi_kernel_W-0.25	te_kraskov_NN-4_k-max-10_tau-max-4
te_kraskov_NN-4_DCE_k-max-10_tau-max-4	te_kraskov_NN-4_DCE_k-2_kt-1_l-1_lt-1	te_kraskov_NN-4_DCE_k-1_kt-1_l-1_lt-1	te_kraskov_NN-4_k-1_kt-1_l-1_lt-1
phi_star_t-1_norm-0	phi_star_t-1_norm-1	phi_Geo_t-1_norm-0	phi_Geo_t-1_norm-1
dtf_multitaper_mean_fs-1_fmin-0_fmax-0-5	dtf_multitaper_mean_fs-1_fmin-0_fmax-0-25	dtf_multitaper_mean_fs-1_fmin-0-25_fmax-0-5	dtf_multitaper_max_fs-1_fmin-0_fmax-0-5
dtf_multitaper_max_fs-1_fmin-0_fmax-0-25	dtf_multitaper_max_fs-1_fmin-0-25_fmax-0-5	dcoh_multitaper_mean_fs-1_fmin-0_fmax-0-5	dcoh_multitaper_mean_fs-1_fmin-0_fmax-0-25
dcoh_multitaper_mean_fs-1_fmin-0-25_fmax-0-5	dcoh_multitaper_max_fs-1_fmin-0_fmax-0-5	dcoh_multitaper_max_fs-1_fmin-0_fmax-0-25	dcoh_multitaper_max_fs-1_fmin-0-25_fmax-0-5
pdcoh_multitaper_mean_fs-1_fmin-0_fmax-0-5	pdcoh_multitaper_mean_fs-1_fmin-0_fmax-0-25	pdcoh_multitaper_mean_fs-1_fmin-0-25_fmax-0-5	pdcoh_multitaper_max_fs-1_fmin-0_fmax-0-5
pdcoh_multitaper_max_fs-1_fmin-0_fmax-0-25	pdcoh_multitaper_max_fs-1_fmin-0-25_fmax-0-5	gpdcoh_multitaper_mean_fs-1_fmin-0_fmax-0-5	gpdcoh_multitaper_mean_fs-1_fmin-0_fmax-0-25
gpdcoh_multitaper_mean_fs-1_fmin-0-25_fmax-0-5	gpdcoh_multitaper_max_fs-1_fmin-0_fmax-0-5	gpdcoh_multitaper_max_fs-1_fmin-0_fmax-0-25	gpdcoh_multitaper_max_fs-1_fmin-0-25_fmax-0-5
ddtf_multitaper_mean_fs-1_fmin-0_fmax-0-5	ddtf_multitaper_mean_fs-1_fmin-0_fmax-0-25	ddtf_multitaper_mean_fs-1_fmin-0-25_fmax-0-5	ddtf_multitaper_max_fs-1_fmin-0_fmax-0-5
ddtf_multitaper_max_fs-1_fmin-0_fmax-0-25	ddtf_multitaper_max_fs-1_fmin-0-25_fmax-0-5	gd_multitaper_delay_fs-1_fmin-0_fmax-0-25	sgc_nonparametric_mean_fs-1_fmin-0_fmax-0-5
sgc_nonparametric_mean_fs-1_fmin-0_fmax-0-25	sgc_nonparametric_mean_fs-1_fmin-0-25_fmax-0-5	sgc_nonparametric_max_fs-1_fmin-0_fmax-0-5	sgc_nonparametric_max_fs-1_fmin-0_fmax-0-25
sgc_nonparametric_max_fs-1_fmin-0-25_fmax-0-5	sgc_parametric_mean_fs-1_fmin-0_fmax-0-5_order-None	sgc_parametric_mean_fs-1_fmin-0_fmax-0-25_order-None	sgc_parametric_mean_fs-1_fmin-0-25_fmax-0-5_order-None
sgc_parametric_mean_fs-1_fmin-1e-05_fmax-0-5_order-1	sgc_parametric_mean_fs-1_fmin-0-25_fmax-0-5_order-1	sgc_parametric_mean_fs-1_fmin-1e-05_fmax-0-5_order-20	sgc_parametric_mean_fs-1_fmin-0_fmax-0-25_order-20
sgc_parametric_mean_fs-1_fmin-0-25_fmax-0-5_order-20	sgc_parametric_max_fs-1_fmin-1e-05_fmax-0-5_order-None	sgc_parametric_max_fs-1_fmin-0_fmax-0-25_order-None	sgc_parametric_max_fs-1_fmin-0-25_fmax-0-5_order-None
sgc_parametric_max_fs-1_fmin-0_fmax-0-25_order-1	sgc_parametric_max_fs-1_fmin-0-25_fmax-0-5_order-1	sgc_parametric_max_fs-1_fmin-1e-05_fmax-0-5_order-20	sgc_parametric_max_fs-1_fmin-0_fmax-0-25_order-20
sgc_parametric_max_fs-1_fmin-0-25_fmax-0-5_order-20	psi_wavelet_mean_fs-1_fmin-0_fmax-0-5_mean	lmfit_Ridge	lmfit_Lasso
lmfit_SGDRegressor	lmfit_ElasticNet	lmfit_BayesianRidge	gpfit_DotProduct
gpfit_RBF	coint_johansen_max_eig_stat_order-0_ardiff-10	coint_johansen_trace_stat_order-0_ardiff-10	coint_johansen_max_eig_stat_order-0_ardiff-1
coint_johansen_trace_stat_order-0_ardiff-1	coint_johansen_max_eig_stat_order-1_ardiff-10	coint_johansen_trace_stat_order-1_ardiff-10	coint_johansen_max_eig_stat_order-1_ardiff-1
coint_johansen_trace_stat_order-1_ardiff-1	coint_aeg_tstat_trend-c_autolag-aic_maxlag-10	coint_aeg_tstat_trend-ct_autolag-aic_maxlag-10	coint_aeg_tstat_trend-ct_autolag-bic_maxlag-10

H Computational Complexity Analysis

H.1 Computational Complexity and Selection of SPIs

We compute all Statistical Pairwise Interactions (SPIs) using the open-source Python library `pyspi` (Cliff et al., 2023), which provides a unified implementation of 284 diverse measures. To ensure numerical stability, we exclude 18 SPIs that produced invalid matrix entries (`NaN` values), resulting in a final set of $|\mathcal{S}| = 266$ methods for our benchmark.

For a single multivariate time series (MTS) sample $\mathbf{X} \in \mathbb{R}^{N \times T}$, where N is the number of regions of interest (ROIs) and T is the number of time points, the time complexity of a typical SPI computation scales as $\mathcal{O}(N^2 L)$, where L reflects the method-specific internal dependency length. When considering the entire benchmarking task over a dataset $\mathcal{X}$ and the full suite of SPIs $\mathcal{S}$, the overall computational complexity lower bound becomes $\mathcal{O}(|\mathcal{X}| \cdot |\mathcal{S}| \cdot N^2 L)$.

To quantify this theoretical burden in practical terms, we benchmarked each of the 266 SPIs on a representative sample (size 33×240). Using a 128 vCPU cluster as a concrete example, the resulting time distribution is shown in Figure A1. Based on these timings, we can estimate the total cost for our full dataset of $|\mathcal{X}| = 4520$ samples. The time to process one sample with all 266 SPIs is approximately $18,950$ CPU-seconds (summing estimates from different time bins: 14 methods $\times$ $1000s + 45 \times 100s + 27 \times 10s + 180 \times 1s$). Extrapolating to the full dataset, the total computational cost is a staggering $4520 \times 18,950 \approx 8.57 \times 10^7$ CPU-seconds, equivalent to over **990 CPU-days** ($\approx$ 7.7 CPU-days on a 128 vCPU). Even with access to massively parallelized cluster environments, this enormous consumption of resources makes a full benchmark practically infeasible and time-prohibitive.

This severe computational bottleneck motivates our core research question: how can we drastically reduce the number of samples while preserving a robust and reliable evaluation of the SPIs? This challenge naturally leads to our investigation of core-set selection for fMRI-based SPI benchmarking. As calculated, on a 10% core-set the load reduces to $\approx$ 99 CPU-days (10% of the full cost). On the same 128-vCPU cluster, this would take: 99 CPU-days / 128 cores = $\approx$ 0.77 days (or $\approx$ 18.5 hours). To validate our approach, we necessarily performed this exhaustive computation to establish a ground-truth ranking. However, for the purpose of evaluating core-set quality in our experiments, we restrict our analysis to a tractable subset of SPIs that take less than one second per sample, enabling rapid yet informative evaluation. The full list of 284 SPIs is summarized in Table A1, with those used for our core-set evaluation highlighted in bold.

H.2 Computational Cost of Selection Methods

To complete our analysis of efficiency, we provide an empirical comparison of the computational cost for **SCLCS** and the baseline methods. Table A2 details the practical time consumption required for each method. The 'Time per Epoch' reflects the average wall-clock time to complete a single training epoch. The 'Score Calculation Time' is the specific, one-time overhead for computing the final selection metric after the training phase is complete.

Table A2: Computational cost for core-set selection methods.

Method	**Time per Epoch (s)**	**Score Calculation Time (s)**
Forgetting	1.8708	0.0000
Entropy	1.9851	2.1588
EL2N	2.5309	0.3420
AUM	2.8389	0.0957
CCS	2.8389	11.9414
EVA	2.5964	0.6479
BOSS	2.0952	61.5406
SCLCS$_{\text{Dense}}$	6.9296	249.8204

As the results indicate, all core-set selection methods incur a modest, one-time computational cost. This up-front investment is negligible when contrasted with the over 990 CPU-days required for a full downstream benchmark (as detailed in **Appendix H**). It is worth noting that the score calculation for **SCLCS**$_{\text{Dense}}$, which computes the **SPS** metric by measuring differences between attention matrices across epochs, represents a fixed, post-training overhead. While this step appears slower than other methods' scoring, it is a one-time process that is highly parallelizable. These findings confirm that investing a small computational budget in core-set selection is a practical and efficient solution. This validates our proposed paradigm for tackling the intractable problem of large-scale model benchmarking.

I Influence of Labels on Baselines

In the main paper, we report results of baseline methods trained using labels aligned with the evaluation objective. However, our proposed **SCLCS** framework uniformly adopts subject identity as the supervisory signal, which raises the question of training-evaluation label misalignment. To provide a more comprehensive analysis, this section further investigates the influence of such misalignment on baselines.

I.1 Using MDD Label for Brain Fingerprinting Ranking

Table A3: Comparison of brain fingerprinting ranking with subject and with MDD labels (mean ± std). Arrows indicate change under MDD supervision.

Method	Label	nDCG@5			nDCG@10			nDCG@20		
		0.1	0.3	0.5	0.1	0.3	0.5	0.1	0.3	0.5
Forgetting	Subject	$14.41_{\pm 7.84}$	$54.43_{\pm 13.26}$	$43.36_{\pm 3.35}$	$15.49_{\pm 7.83}$	$48.57_{\pm 7.85}$	$49.87_{\pm 4.66}$	$20.23_{\pm 7.21}$	$55.80_{\pm 6.72}$	$49.70_{\pm 4.60}$
	MDD	$14.41_{\pm 7.84}\downarrow$	$54.43_{\pm 13.26}\downarrow$	$43.36_{\pm 3.34}\uparrow$	$15.49_{\pm 7.83}\downarrow$	$48.57_{\pm 7.85}\downarrow$	$49.87_{\pm 4.66}\downarrow$	$20.23_{\pm 7.21}\downarrow$	$55.80_{\pm 6.72}\downarrow$	$49.70_{\pm 4.60}\downarrow$
Entropy	Subject	$47.40_{\pm 40.26}$	$22.84_{\pm 13.20}$	$59.95_{\pm 26.99}$	$37.73_{\pm 25.94}$	$32.05_{\pm 15.22}$	$58.68_{\pm 23.71}$	$36.05_{\pm 18.11}$	$35.90_{\pm 16.19}$	$57.72_{\pm 21.84}$
	MDD	$24.38_{\pm 23.28}\downarrow$	$36.60_{\pm 9.47}\uparrow$	$68.74_{\pm 19.96}\uparrow$	$29.45_{\pm 28.94}\downarrow$	$44.86_{\pm 25.46}\uparrow$	$58.27_{\pm 13.07}\downarrow$	$33.17_{\pm 25.94}\downarrow$	$46.55_{\pm 13.74}\uparrow$	$59.05_{\pm 11.16}\uparrow$
El2N	Subject	$35.56_{\pm 36.16}$	$20.30_{\pm 5.77}$	$31.03_{\pm 34.57}$	$36.51_{\pm 25.38}$	$33.18_{\pm 14.68}$	$32.70_{\pm 30.84}$	$33.30_{\pm 21.55}$	$35.82_{\pm 12.10}$	$40.06_{\pm 32.08}$
	MDD	$4.51_{\pm 4.67}\downarrow$	$19.27_{\pm 15.05}\downarrow$	$44.09_{\pm 25.52}\uparrow$	$14.67_{\pm 20.47}\downarrow$	$22.51_{\pm 11.13}\downarrow$	$47.17_{\pm 20.08}\uparrow$	$14.60_{\pm 18.20}\downarrow$	$24.27_{\pm 10.47}\downarrow$	$48.58_{\pm 17.72}\uparrow$
AUM	Subject	$65.92_{\pm 33.80}$	$56.68_{\pm 11.11}$	$38.17_{\pm 13.94}$	$60.95_{\pm 30.91}$	$62.05_{\pm 4.91}$	$36.83_{\pm 8.78}$	$51.75_{\pm 22.42}$	$59.09_{\pm 5.72}$	$38.34_{\pm 6.88}$
	MDD	$42.32_{\pm 42.32}\downarrow$	$72.58_{\pm 23.18}\uparrow$	$48.75_{\pm 30.04}\uparrow$	$36.06_{\pm 29.57}\downarrow$	$63.74_{\pm 19.11}\uparrow$	$53.60_{\pm 25.75}\uparrow$	$36.70_{\pm 19.54}\downarrow$	$65.14_{\pm 17.92}\uparrow$	$53.85_{\pm 22.38}\uparrow$
CCS	Subject	$1.90_{\pm 2.07}$	$30.53_{\pm 14.15}$	$46.65_{\pm 22.32}$	$2.92_{\pm 3.24}$	$29.13_{\pm 8.00}$	$51.78_{\pm 24.12}$	$16.24_{\pm 13.34}$	$32.56_{\pm 14.15}$	$52.18_{\pm 20.20}$
	MDD	$11.75_{\pm 11.52}\uparrow$	$41.01_{\pm 24.51}\uparrow$	$24.68_{\pm 20.39}\downarrow$	$12.93_{\pm 12.11}\uparrow$	$44.33_{\pm 21.58}\uparrow$	$33.11_{\pm 26.79}\downarrow$	$18.45_{\pm 10.26}\uparrow$	$44.20_{\pm 22.46}\uparrow$	$33.65_{\pm 26.74}\downarrow$
EVA	Subject	$38.40_{\pm 40.57}$	$62.03_{\pm 26.64}$	$37.80_{\pm 42.45}$	$43.37_{\pm 19.92}$	$55.01_{\pm 20.05}$	$49.56_{\pm 33.28}$	$43.22_{\pm 15.28}$	$53.51_{\pm 14.70}$	$65.49_{\pm 21.99}$
	MDD	$31.99_{\pm 32.72}\downarrow$	$30.57_{\pm 29.76}\downarrow$	$42.34_{\pm 42.47}\uparrow$	$30.85_{\pm 30.00}\downarrow$	$31.90_{\pm 30.09}\downarrow$	$42.55_{\pm 25.48}\downarrow$	$41.33_{\pm 15.02}\downarrow$	$28.07_{\pm 23.41}\downarrow$	$53.45_{\pm 16.83}\downarrow$
BOSS	Subject	$15.98_{\pm 25.58}$	$42.11_{\pm 24.33}$	$35.36_{\pm 9.21}$	$29.44_{\pm 11.05}$	$40.57_{\pm 23.37}$	$39.15_{\pm 6.71}$	$31.45_{\pm 9.45}$	$38.24_{\pm 19.65}$	$38.92_{\pm 5.97}$
	MDD	$22.90_{\pm 23.99}\uparrow$	$51.70_{\pm 13.14}\uparrow$	$40.08_{\pm 31.21}\uparrow$	$22.92_{\pm 18.55}\downarrow$	$51.87_{\pm 12.74}\uparrow$	$39.62_{\pm 29.12}\uparrow$	$29.20_{\pm 17.36}\downarrow$	$51.71_{\pm 10.03}\uparrow$	$40.14_{\pm 28.44}\uparrow$

We first evaluate the performance of baseline methods trained with MDD diagnosis labels for core-set selection in the brain fingerprinting ranking task, using the results from subject-identity supervision (as reported in the main paper) as reference. This alternative labeling scheme aligns with the intended design of most baseline algorithms, which aim to select core-sets based on the target task labels (i.e., MDD vs. HC diagnosis labels from the REST-meta-MDD dataset).

As shown in Table A3, this supervision shift generally leads to performance degradation across most methods, suggesting a potential mismatch between the binary nature of MDD labels and the requirements of brain fingerprinting, which involves a one-vs-all subject-level identification task. Specifically, training with MDD labels provides weaker sample-level supervision due to reduced class granularity, potentially limiting the diversity captured during core-set selection. Consequently, the selected samples may fail to adequately support the subject-wise discriminative capacity of SPIs.

Nevertheless, certain methods such as AUM and BOSS exhibit improved ranking stability at higher sampling ratios. This may be attributed to their scoring strategies, which explicitly account for sample diversity or informativeness. Under this paradigm, incorporating diagnosis-based labels introduces an additional semantic dimension that enhances the selection process, enabling these methods to better preserve the representational structure of the full dataset.

I.2 Using Subject Label for MDD Ranking

As a complementary analysis, we also evaluate the performance of baselines trained with subject identity labels for core-set selection in the MDD diagnosis ranking task. This setting closely aligns with our proposed SCLCS framework and allows us to investigate whether individual-level supervision provides a stronger basis for core-set construction.

Table A4: Comparison of MDD diagnosis ranking with MDD labels and with subject identities(mean ± std). Arrows indicate change under subject supervision.

Method	Label	nDCG@5			nDCG@10			nDCG@20		
		0.1	0.3	0.5	0.1	0.3	0.5	0.1	0.3	0.5
Forgetting	MDD	$28.70_{\pm 28.06}$	$41.60_{\pm 28.08}$	$61.42_{\pm 4.31}$	$40.56_{\pm 28.06}$	$50.27_{\pm 32.37}$	$66.72_{\pm 8.60}$	$44.37_{\pm 28.07}$	$57.02_{\pm 33.39}$	$75.70_{\pm 7.58}$
	Subject	$24.64_{\pm 28.64}$↓	$43.69_{\pm 36.57}$↑	$87.74_{\pm 10.67}$↑	$28.52_{\pm 30.48}$↓	$46.33_{\pm 37.31}$↓	$88.57_{\pm 11.93}$↑	$32.88_{\pm 32.35}$↓	$51.60_{\pm 38.74}$↓	$90.93_{\pm 8.89}$↑
Entropy	MDD	$29.00_{\pm 45.10}$	$31.84_{\pm 46.19}$	$57.48_{\pm 29.81}$	$29.17_{\pm 43.24}$	$40.75_{\pm 44.05}$	$63.23_{\pm 23.44}$	$30.30_{\pm 43.52}$	$45.85_{\pm 45.00}$	$70.79_{\pm 18.00}$
	Subject	$53.09_{\pm 17.82}$↑	$46.99_{\pm 34.99}$↑	$66.46_{\pm 19.72}$↑	$58.79_{\pm 16.78}$↑	$52.25_{\pm 36.91}$↑	$75.19_{\pm 18.27}$↑	$65.36_{\pm 13.54}$↑	$56.61_{\pm 37.02}$↑	$81.40_{\pm 14.68}$↑
El2N	MDD	$30.93_{\pm 41.17}$	$51.70_{\pm 28.14}$	$68.85_{\pm 12.46}$	$36.05_{\pm 24.41}$	$58.54_{\pm 27.55}$	$72.96_{\pm 15.77}$	$41.04_{\pm 39.61}$	$64.68_{\pm 26.25}$	$78.85_{\pm 14.56}$
	Subject	$68.75_{\pm 21.44}$↑	$42.90_{\pm 32.24}$↓	$60.18_{\pm 17.44}$↓	$71.38_{\pm 22.07}$↑	$50.92_{\pm 30.16}$↓	$66.67_{\pm 18.04}$↓	$77.12_{\pm 16.54}$↑	$56.72_{\pm 28.89}$↓	$73.55_{\pm 15.61}$↓
AUM	MDD	$36.07_{\pm 12.59}$	$42.68_{\pm 46.16}$	$35.49_{\pm 14.37}$	$39.17_{\pm 14.58}$	$44.41_{\pm 44.05}$	$58.14_{\pm 16.54}$	$44.94_{\pm 18.48}$	$48.44_{\pm 42.48}$	$64.36_{\pm 14.40}$
	Subject	$41.59_{\pm 10.89}$↑	$52.72_{\pm 39.35}$↑	$65.20_{\pm 46.67}$↑	$53.62_{\pm 5.24}$↑	$58.66_{\pm 37.39}$↑	$67.17_{\pm 44.70}$↑	$53.62_{\pm 5.24}$↑	$65.26_{\pm 36.40}$↑	$70.07_{\pm 41.78}$↑
CCS	MDD	$55.95_{\pm 16.28}$	$59.92_{\pm 30.08}$	$74.73_{\pm 12.46}$	$59.25_{\pm 16.26}$	$67.01_{\pm 24.61}$	$75.41_{\pm 18.08}$	$68.98_{\pm 15.67}$	$71.62_{\pm 21.15}$	$78.77_{\pm 13.49}$
	Subject	$47.81_{\pm 27.72}$↓	$25.25_{\pm 8.66}$↓	$56.95_{\pm 27.90}$↓	$55.46_{\pm 21.63}$↓	$32.40_{\pm 8.73}$↓	$62.59_{\pm 25.19}$↓	$62.04_{\pm 17.97}$↓	$41.36_{\pm 7.82}$↓	$69.14_{\pm 19.28}$↓
EVA	MDD	$31.80_{\pm 17.70}$	$66.81_{\pm 9.12}$	$70.07_{\pm 6.05}$	$36.11_{\pm 14.79}$	$72.61_{\pm 8.73}$	$76.26_{\pm 6.17}$	$48.39_{\pm 19.48}$	$75.34_{\pm 8.48}$	$81.73_{\pm 5.61}$
	Subject	$21.25_{\pm 11.48}$↓	$67.32_{\pm 30.66}$↑	$75.63_{\pm 16.84}$↑	$28.40_{\pm 10.75}$↓	$73.84_{\pm 21.80}$↑	$79.06_{\pm 15.46}$↑	$36.98_{\pm 10.08}$↓	$80.82_{\pm 15.48}$↑	$83.18_{\pm 12.42}$↑
BOSS	MDD	$42.44_{\pm 24.37}$	$57.52_{\pm 20.11}$	$79.57_{\pm 16.62}$	$50.64_{\pm 25.12}$	$64.86_{\pm 21.86}$	$84.70_{\pm 14.07}$	$58.11_{\pm 23.22}$	$71.36_{\pm 18.94}$	$88.95_{\pm 9.49}$
	Subject	$39.60_{\pm 37.05}$↓	$67.40_{\pm 13.42}$↑	$79.51_{\pm 17.82}$↓	$45.30_{\pm 31.25}$↓	$68.47_{\pm 15.58}$↑	$86.21_{\pm 9.45}$↑	$54.29_{\pm 2.51}$↓	$77.59_{\pm 11.02}$↑	$87.29_{\pm 8.34}$↓

Compared to the MDD-supervised setting, we observe that a greater number of methods benefit from improved ranking stability under subject identity supervision. For instance, Entropy, AUM, and El2N exhibit consistent performance gains at higher sampling ratios (e.g., AUM improves from 58.14% to 67.17% in nDCG@10@0.5, and Entropy from 63.23% to 75.19%). This trend suggests that supervision aligned with subject-level heterogeneity may better preserve fine-grained information necessary for identifying high-quality representative samples.

Despite these improvements, a clear performance gap remains between all baselines and our method, as shown in Table 3. This highlights the non-trivial advantage of **SCLCS**, where structural perturbation scoring and density-aware sampling jointly enforce both informativeness and diversity in selected subsets.

Interestingly, AUM shows consistent gains across nearly all metrics, suggesting that its original scoring, formulated under task-specific supervision, may underestimate structural variation among samples. Subject-based training appears to compensate for this limitation by injecting more diverse contrastive signals during scoring, revealing a potential direction for enhancing its robustness.

Entropy also achieves performance gains across the board, with improvements as high as 11.6% in nDCG@20 (from 70.79% to 81.40%). However, as shown in Figure 2, Entropy consistently selects highly imbalanced subsets regardless of supervision label, often dominated by a small number of subjects. This structural bias undermines its utility for benchmarking core-set methods, as it fails to preserve a representative distribution of the dataset. Thus, despite the numerical improvements, Entropy remains unsuitable for core-set-based SPI evaluation.

These findings reinforce the need to consider both label alignment and structural diversity in core-set selection. Subject-level supervision offers a promising direction, but our method's explicit modeling of structure-aware consistency and coverage remains critical for reliable benchmarking.

J Influence of the Proposed Density-balanced Sampling on Baselines

As stated in **Section 4.3**, our proposed density-based sampling strategy is not only central to the **SCLCS** framework, but also generalizable to other score-based core-set selection methods. To evaluate its applicability beyond our method, we visualize comparative results on the Brain Fingerprinting and MDD Diagnosis tasks in Figure A2 and Figure A3, respectively. We exclude CCS and BOSS from this analysis due to their use of task-specific sampling designs.

On the Brain Fingerprinting task, density-based sampling frequently alters SPI ranking performance across baselines. For example, El2N underperforms in multiple metrics when density is applied (e.g., nDCG@5 at ratio 0.1 drops from ∼40 to ∼20), while EVA shows inconsistent trends across sampling ratios. This instability may be attributed to the fact that, when supervised with subject identity, score-based methods tend to prioritize samples that support subject-level diversity. Replacing this priority with a density-based criterion may inadvertently distort the structural balance of the selected subset.

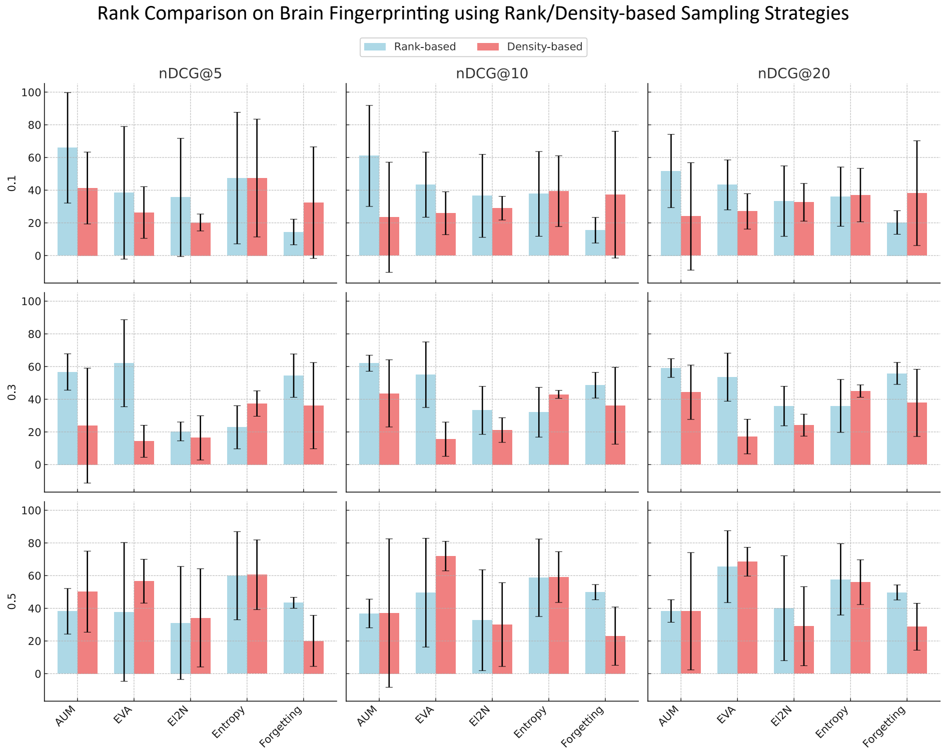

Figure A2: Rank comparison on brain fingerprinting using rank/density-based sampling strategies.

Conversely, in the MDD Diagnosis setting, density-based sampling consistently improves performance. Baselines such as EVA and Entropy show marked gains in ranking stability, with EVA's nDCG@10 increasing from ∼70 to ∼90 at sampling ratio 0.5. This contrast suggests that when supervision involves fewer discrete classes (e.g., binary labels), the density structure becomes easier to estimate and more semantically aligned with the downstream evaluation objective.

An interesting deviation is observed in the Forgetting method. At high sampling ratios in both tasks, its performance noticeably declines under density-based sampling. This may stem from Forgetting's underlying assumption: that low-confidence samples correspond to noisy or uninformative data, which does not hold well for fMRI. Due to the complex nature of brain dynamics, such samples may actually be densely clustered and structurally meaningful. Consequently, density-based sampling could overemphasize regions marked by high forgetting scores, thereby degrading ranking consistency.

In summary, density-based sampling demonstrates strong compatibility with several baseline strategies, highlighting its potential as a general-purpose augmentation. However, the interaction between density criteria and different scoring heuristics can be nontrivial, and may lead to unintended trade-offs in performance. These results underscore the importance of further empirical studies to better understand the conditions determining whether density-based selection benefits or interferes with score-driven core-set construction.

K Empirical Results for Theorems

K.1 Convergence of Universal Approximation

Experimental Setup To empirically validate the universal approximation capability of our modified Transformer architecture, as posited in **Theorem 2**, we designed and conducted a direct fitting experiment. The theorem states that our model architecture is, in principle, capable of approximating any continuous Statistical Pairwise Interaction (SPI) operator.

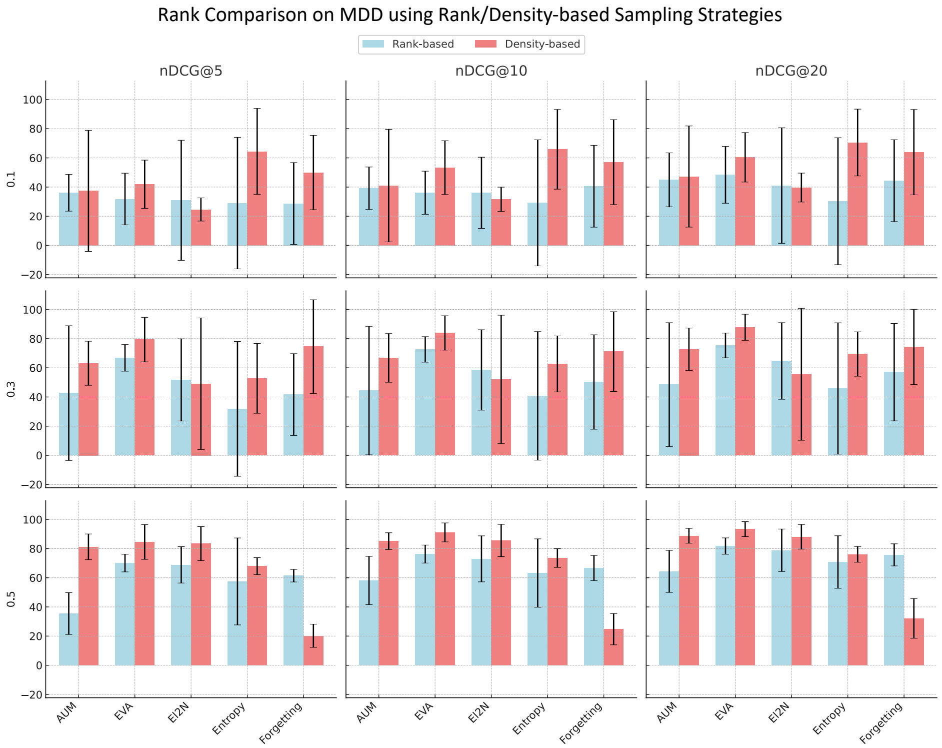

Figure A3: Rank comparison on MDD diagnosis using rank/density-based sampling strategies.

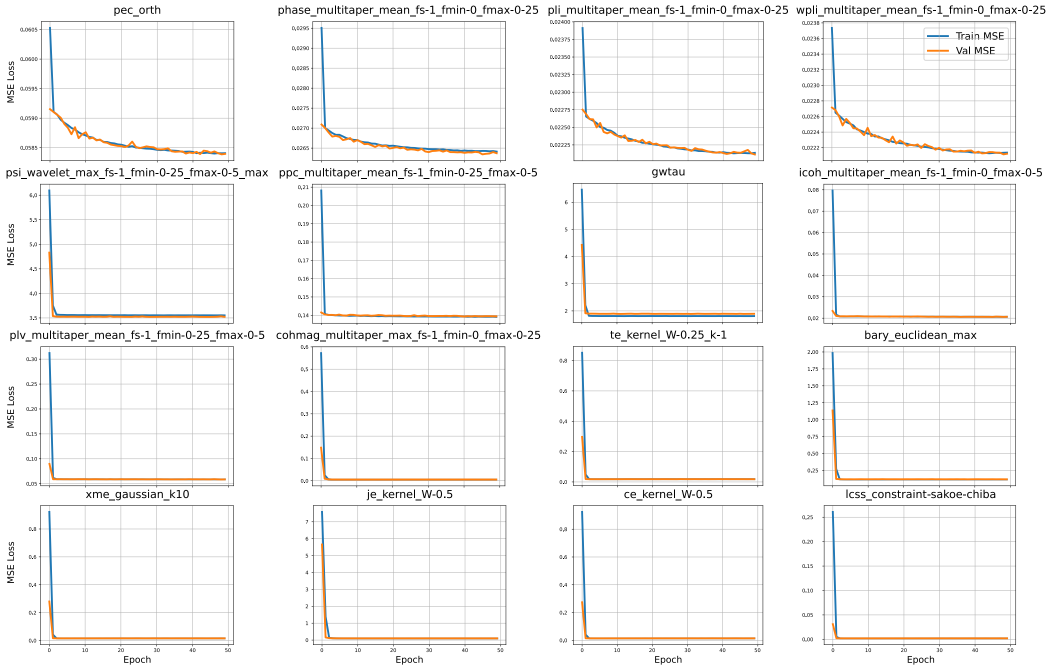

Figure A4: Training and validation MSE loss convergence curves for the 16 SPI operators used in the empirical validation of **Theorem 2**.

The core objective of this experiment was to train instances of our model to directly mimic the Functional Connectivity (FC) matrices produced by specific SPI operators. We trained a separate model instance for each operator in a diverse set of 16 representative SPIs. The experimental setup was as follows:

- **Input**: The input for each model was an fMRI time series sample $\mathbf{X} \in \mathbb{R}^{N \times T}$, consistent with the primary data used in our paper.
- **Target**: For each model, the learning target was the "ground-truth" FC matrix $\mathbf{A}_S \in \mathbb{R}^{N \times N}$, computed for that specific fMRI sample using the corresponding SPI operator.
- **Loss Function**: We optimized the model parameters by directly minimizing the Mean Squared Error (MSE) between the model's output adjacency matrix $\mathbf{A}$ and the target matrix $\mathbf{A}_S$. Specifically, we use $\mathrm{MSE}((\mathbf{A} + \mathbf{A}^T)/2, (\mathbf{A}_S + \mathbf{A}_S^T)/2)$ to ensure the symmetry of the adjacencies.
- **Training and Validation**: We partitioned the dataset into training, validation, and testing sets with a 70%/10%/20% split. During training, we employed early stopping (patience=10) based on the validation set performance to ensure the model learned a generalizable transformation rather than merely overfitting to the training data.

Results and Analysis Figure A4 displays the convergence curves from the training process for all 16 SPI operators. Each subplot in the figure represents an independent model instance, with the x-axis denoting the training epoch and the y-axis representing the MSE Loss. The blue line indicates the MSE on the training set, while the orange line represents the MSE on the validation set.

As the plots clearly demonstrate, our model architecture exhibits a strong capacity to fit all 16 SPIs, regardless of their diverse underlying computational principles. Key observations include:

1. **Successful Convergence**: In all experiments, both the training and validation MSE decrease rapidly from a high initial value and eventually converge to a stable, low level. This indicates that the optimization process was successful and that the model effectively learned the mapping from the fMRI time series to the target FC matrix.
2. **Good Generalization**: The validation loss curves closely track the training loss curves without significant divergence. This confirms that the models did not overfit and that the learned approximations generalize well to unseen data.

These convergence curves, combined with the low final test MSE values reported in Table 4 of the main text, provide strong empirical support for **Theorem 2**. The results collectively confirm that our proposed modified Transformer architecture possesses the practical expressive power required to represent the diverse functional forms inherent to our benchmarking task.

K.2 Stationary and Convergence Analysis

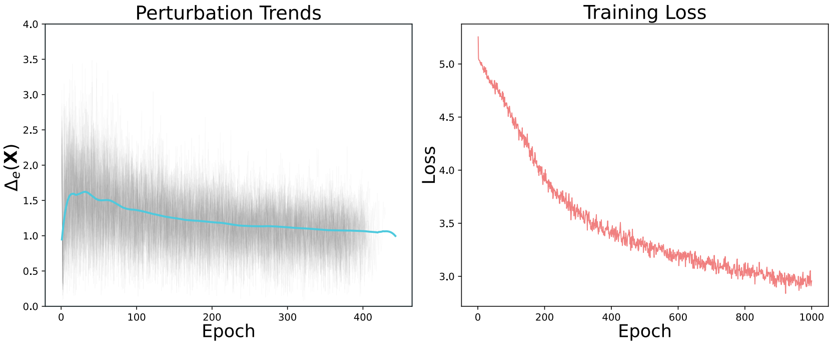

Figure A5: Convergence dynamics of **SCLCS**. **(Left)** Perturbation trends show that the mean **SPS** stabilizes relatively early in training. **(Right)** The training loss converges more gradually over 1000 epochs. The different x-axes are used to visualize the distinct convergence timescales of each metric.

As demonstrated in **Lemma 1**, the reliability of the Structural Perturbation Score (**SPS**) relies on the assumption that the per-epoch difference $\Delta_e(\mathbf{X}) = \|\mathbf{A}_{(\mathbf{X})}^{(e)} - \mathbf{A}_{(\mathbf{X})}^{(e-1)}\|_F^2$ is stationary and ergodic.

Table A5: Performance comparison: brain fingerprinting vs. MDD diagnosis

Method	**Brain Fingerprinting**									**MDD Diagnosis**								
	nDCG@5			nDCG@10			nDCG@20			nDCG@5			nDCG@10			nDCG@20		
Ratio	0.1	0.3	0.5	0.1	0.3	0.5	0.1	0.3	0.5	0.1	0.3	0.5	0.1	0.3	0.5	0.1	0.3	0.5
SECTION A: Slow-Only SPIs																		
Random	0.883	0.863	0.805	0.903	0.898	0.835	0.921	0.923	0.863	0.946	0.998	0.997	0.962	0.998	0.998	0.975	0.999	0.998
Forgetting	0.591	0.441	0.837	0.614	0.478	0.871	0.642	0.514	0.898	0.781	0.946	0.997	0.839	0.965	0.998	0.886	0.977	0.999
Entropy	0.544	0.963	0.907	0.615	0.970	0.977	0.683	0.975	0.982	0.604	0.636	0.941	0.619	0.640	0.959	0.628	0.652	0.973
EI2N	0.887	0.852	0.928	0.912	0.884	0.945	0.928	0.909	0.958	0.967	0.897	0.994	0.976	0.930	0.996	0.983	0.957	0.997
AUM	0.835	0.983	0.808	0.866	0.984	0.848	0.905	0.983	0.881	0.934	0.849	0.933	0.953	0.890	0.955	0.963	0.923	0.971
CCS	0.785	0.737	0.845	0.822	0.790	0.880	0.859	0.832	0.908	0.840	0.966	0.999	0.887	0.978	0.999	0.923	0.985	0.999
EVA	0.628	0.906	0.650	0.663	0.923	0.707	0.701	0.935	0.761	0.795	0.997	0.999	0.848	0.998	0.999	0.895	0.999	0.999
BOSS	0.385	0.956	0.602	0.444	0.951	0.656	0.521	0.945	0.710	0.636	0.968	0.992	0.760	0.976	0.995	0.843	0.988	0.995
SCLCS	0.879	0.941	0.999	0.907	0.955	0.999	0.932	0.961	0.999	0.874	0.983	0.997	0.916	0.989	0.998	0.945	0.993	0.999
SCLCS$_{\text{Dense}}$	0.903	0.962	0.943	0.922	0.967	0.956	0.935	0.969	0.964	0.954	0.999	0.999	0.969	0.999	0.999	0.979	0.999	0.999
SECTION B: Mixture SPIs																		
Random	0.280	0.680	0.646	0.389	0.689	0.715	0.392	0.765	0.722	0.504	0.242	0.718	0.601	0.263	0.779	0.633	0.306	0.828
Forgetting	0.323	0.464	0.711	0.421	0.485	0.742	0.465	0.497	0.749	0.255	0.415	0.634	0.380	0.502	0.670	0.432	0.563	0.758
Entropy	0.442	0.660	0.663	0.485	0.656	0.678	0.551	0.668	0.684	0.289	0.327	0.565	0.291	0.414	0.621	0.307	0.467	0.702
EI2N	0.485	0.441	0.531	0.552	0.515	0.570	0.564	0.570	0.623	0.310	0.522	0.656	0.360	0.576	0.705	0.416	0.646	0.761
AUM	0.491	0.680	0.870	0.565	0.737	0.882	0.604	0.767	0.877	0.361	0.415	0.536	0.403	0.435	0.585	0.452	0.478	0.640
CCS	0.228	0.480	0.495	0.345	0.478	0.507	0.352	0.576	0.549	0.469	0.599	0.747	0.519	0.670	0.754	0.628	0.699	0.790
EVA	0.193	0.527	0.601	0.362	0.597	0.622	0.396	0.640	0.655	0.274	0.529	0.688	0.353	0.570	0.755	0.440	0.616	0.812
BOSS	0.438	0.592	0.505	0.433	0.662	0.531	0.428	0.687	0.561	0.373	0.620	0.796	0.435	0.620	0.859	0.519	0.721	0.879
SCLCS	0.790	0.704	0.785	0.795	0.734	0.807	0.823	0.755	0.819	0.279	0.703	0.631	0.303	0.708	0.613	0.368	0.764	0.680
SCLCS$_{\text{Dense}}$	0.442	0.758	0.812	0.552	0.751	0.827	0.576	0.750	0.827	0.447	0.449	0.716	0.522	0.533	0.706	0.621	0.594	0.752

Since our training does not incorporate validation-based early stopping, we use model convergence as an implicit criterion for termination. To empirically support this, Figure A5 visualizes the convergence trends of structural perturbation and training loss.

Notably, these two metrics converge on different effective timescales. The left panel shows that the mean **SPS** (blue curve), which reflects the stability of the learned connectivity structures, reaches a stable plateau relatively early (around epoch 500). The initial spike in perturbation reflects an expected "burn-in" phase before the model learns stable representations. In contrast, the right panel shows that the training loss continues to decrease more gradually over the full 1000 epochs as the model makes finer adjustments to the embedding space to fully optimize the contrastive objective.

The clear convergence of both metrics, despite their different timescales, provides strong empirical validation for the assumptions in **Lemma 1**. This confirms that **SPS** is a stable and consistent measure of structural influence, and the overall loss convergence supports the robustness of our end-to-end training pipeline for core-set selection.

L Generalization and Robustness Analysis

This section provides experiment results to address the concerns regarding site-level generalization and SPI subset bias to prove the impact of the proposed task.

L.1 Extension on SPI Properties

We create two new benchmark sets: (1) A "Slower-Only" set containing 50 SPIs with $>$ 1s compute time (up to 10s), and (2) A "Mixture" set combining these 50 slow SPIs with our original 130 fast SPIs (Total = 180). Results are reported in Table A5.

The findings are two-fold:

- On "Slow-Only": Surprisingly, the ranking task becomes "easier." Even the Random baseline performs robustly (e.g., nDCG@5 > 0.9 on the MDD task). This suggests that for this specific subset of slower methods, performance is less sensitive to specific data selection.
- On "Mixture" (Full Benchmark): However, when combining fast and slow **SPIs**, **SCLCS** re-emerges as the clear winner. On the Brain Fingerprinting task (10% ratio), **SCLCS** achieves an nDCG@5 of 0.79, whereas Random collapses to 0.28.

It proves that simple heuristics (like Random) are brittle and fail in realistic, heterogeneous benchmarking scenarios (the "Mixture" case). **SCLCS** is necessary because it robustly handles the full spectrum of SPI behaviors, effectively weighting the top-performing methods that researchers care about most. More importantly, it suggests that different properties of the candidate models may affect the difficulty of the proposed problem. Exploring such effect is a promising future direction.

Table A6: Site balance on core-set (L1 distance, lower better)

Method	**L1 Distance @ 0.1 ratio**		**L1 Distance @ 0.3 ratio**		**L1 Distance @ 0.5 ratio**	
	Sample Level	Subject Level	Sample Level	Subject Level	Sample Level	Subject Level
Forgetting	0.239	0.243	0.184	0.107	0.126	0.048
Entropy	0.186	0.139	0.149	0.112	0.146	0.157
EI2N	0.303	0.350	0.260	0.220	0.235	0.195
AUM	0.157	0.143	0.094	0.181	0.057	0.176
CCS	0.202	0.183	0.143	0.175	0.103	0.148
EVA	0.111	0.101	0.047	0.017	0.024	0.081
BOSS	0.737	0.362	0.457	0.107	0.248	0.022
SCLCS	0.127	0.146	0.096	0.039	0.068	0.011
SCLCS$_{\text{Dense}}$	0.195	0.146	0.112	0.054	0.071	0.017

L.2 Extension on Data Properties

We further explore the relationship between 'site ID' and the core-sets selected by all the baseline methods and the proposed **SCLCS** (which is blind to site). Performances are quantified by calculating the L1 Distance between the site distribution of the original dataset and the site distribution of the selected core-set. The samples of each subject are same on the original dataset but vary on the core-set. Thus we provide two levels of result based on sample and subject, respectively. Results are reported in Table A6.

SCLCS demonstrates robustness to site imbalance. For example, at a 30% sampling ratio, **SCLCS** achieves an L1 distance of ≈ 0.096, ranking second only to the variance-based method EVA (≈ 0.047). Crucially, **SCLCS** significantly outperforms other baselines like Entropy (0.149), which tend to be more biased towards specific sites. This is an interesting finding. It necessitates exploring methods to balance different data properties and the performance on the proposed ranking preservation task, which is beyond the scope of our current paper.

L.3 Robustness to Window Size

We re-process the entire dataset using a different sliding window configuration (Window Size = 50 TRs, Stride = 45 TRs), which differs significantly from the setting used in the main paper (Window Size = 70, Stride = 35). We then re-calculated all 130 SPIs and re-evaluated the ranking consistency of **SCLCS** against all baselines on both downstream tasks.

The results, detailed in Table A7, empirically demonstrate that **SCLCS** maintains its superior performance and ranking stability regardless of the window size, confirming that our method captures intrinsic structural patterns rather than artifacts of specific preprocessing parameters.

On MDD diagnosis task, even with shorter window lengths (which can introduce more noise), **SCLCS** and **SCLCS**$_{\text{Dense}}$ consistently outperform all baselines. Notably, at the challenging 0.1 sampling ratio, **SCLCS**$_{\text{Dense}}$ achieves an nDCG@5 of 0.661, which is significantly higher than the strongest baselines like AUM (0.452) and CCS (0.464). This confirms that our density-aware selection strategy is highly robust to variations in temporal segmentation.

On brain fingerprinting task, **SCLCS** continues to show state-of-the-art performance, particularly at the 0.1 ratio with an nDCG@5 of 0.475, far exceeding Random (0.206). While baselines like El2N show high variance (performing well at 0.3 but dropping significantly at 0.5), **SCLCS** and **SCLCS**$_{\text{Dense}}$ exhibit a more stable performance trajectory as the sampling ratio increases (e.g., **SCLCS** improves from 0.475 to 0.673).

Table A7: Robustness of different methods to window size

Method	**Robustness on MDD Diagnosis Task**									**Robustness on Brain Fingerprinting Task**								
	nDCG@5			nDCG@10			nDCG@20			nDCG@5			nDCG@10			nDCG@20		
Ratio	0.1	0.3	0.5	0.1	0.3	0.5	0.1	0.3	0.5	0.1	0.3	0.5	0.1	0.3	0.5	0.1	0.3	0.5
Random	0.213	0.498	0.802	0.413	0.611	0.843	0.558	0.707	0.886	0.206	0.285	0.282	0.234	0.386	0.353	0.270	0.443	0.376
Forgetting	0.234	0.190	0.637	0.336	0.287	0.689	0.469	0.376	0.745	0.252	0.478	0.417	0.287	0.525	0.461	0.324	0.556	0.500
Entropy	0.258	0.467	0.566	0.293	0.506	0.633	0.360	0.533	0.717	0.227	0.474	0.576	0.243	0.526	0.582	0.347	0.544	0.591
EI2N	0.174	0.007	0.531	0.253	0.199	0.567	0.344	0.250	0.597	0.372	**0.670**	0.333	0.366	**0.638**	0.395	0.374	**0.631**	0.405
AUM	0.452	0.631	0.701	0.490	0.725	0.784	0.528	0.797	0.852	0.291	0.690	0.399	0.324	0.684	0.483	0.332	0.728	0.522
CCS	0.464	0.596	0.727	0.554	0.682	0.794	0.625	0.757	0.853	0.240	0.502	0.473	0.282	0.560	0.506	0.306	0.566	0.523
EVA	0.347	0.422	0.513	0.456	0.496	0.597	0.560	0.580	0.663	0.430	0.456	0.444	0.417	0.510	0.476	0.474	0.517	0.488
BOSS	0.281	0.369	0.833	0.380	0.430	0.867	0.468	0.493	0.899	0.062	0.364	0.669	0.110	0.379	0.747	0.151	0.416	0.752
SCLCS	**0.621**	**0.834**	0.708	**0.662**	**0.818**	0.766	**0.735**	**0.849**	0.826	**0.475**	0.569	0.673	**0.491**	0.569	**0.750**	0.495	0.593	**0.762**
SCLCS$_{\text{Dense}}$	**0.661**	0.732	**0.836**	**0.718**	0.781	**0.885**	**0.778**	0.826	**0.917**	0.420	0.593	**0.712**	0.459	0.617	0.707	**0.496**	0.657	0.705

These empirical results directly demonstrate that **SCLCS** is not sensitive to the choice of window size and can reliably select high-quality core-sets across different pipeline configurations. We believe this evidence strongly supports the practical reliability of **SCLCS** in diverse experimental settings.

M Reproducibility Information

M.1 Data Preprocessing

We use a subset of 904 subjects (458 MDD, 446 controls) from the REST-meta-MDD consortium, selected via a two-stage filtering process. First, we adopt the quality-controlled cohort defined in Yan et al. (2019), which retains $1,642$ subjects across 17 sites. Second, we intersect this cohort with the 9-site subset used in Long et al. (2020), chosen for consistent acquisition parameters (3T scanners, 240 time points, TR = 2000 ms). This yields a final sample from 7 sites with harmonized imaging protocols.

The analysis focuses on 33 regions of interest (ROIs) associated with the default mode network (DMN), as defined in the Dosenbach-160 atlas (Dosenbach et al., 2010). ROI-level time series were obtained directly from the preprocessed data released by REST-meta-MDD. We use the version with global signal regression (GSR) applied, consistent with prior findings (Yan et al., 2019).

Our decision to focus on the DMN-33 ROI was a principled decision based on scientific control, reproducibility, and computational feasibility:

- **Scientific Control:** To validate the proposed new 'ranking preservation' task, our first priority was to isolate variables. Our primary goal is to define the "Ranking Preservation" problem and prove that a "structure-based" approach is a feasible path. Introducing multiple atlases or pipelines would change our scenario from Benchmark (SPIs) to a different, combinatorially explosive scenario of Benchmark (SPIs x atlases/pipelines), which is a valuable, promising but separate scientific task.
- **Reproducibility:** This scientific control principle is reinforced by the dataset itself. To ensure the highest standard of reproducibility, we used the official, standardized pre-processed data from the REST-meta-MDD consortium. This setting ensures the results are not affected by personally defined preprocessing pipelines.
- **The Computational Infeasibility:** Finally, even if we were to use an official atlas like AAL-90 (N = 90), an average complexity of each SPI would increase compute time by ≈ 9x (from N = 33).

Each subject is represented by a multivariate time series of $T \times R$, where T is the number of time points and $R = 33$. To standardize downstream sampling, we truncate all time series to 210 time points. We apply a sliding window of length 70 TRs with a step of 35 TRs, yielding five overlapping temporal segments per subject. This configuration is consistent with prior dynamic connectivity studies (Allen et al., 2014; Preti et al., 2017; Long et al., 2020), and provides a balance between temporal resolution and estimation reliability. These segments serve as dynamic samples for our core-set selection framework, alongside static networks built from each subject's full time series.

Table A8: Acquisition details of the 7 selected REST-meta-MDD sites. All data were collected using 3T scanners and have consistent TR $(2,000$ ms). Minor variation exists in number of time points.

Site ID	Institution	MDD	HC	Scanner	TR (ms)	TE (ms)	Timepoints
15	Zhongda Hospital, Southeast University	37	30	Siemens Verio 3T	2000	25.0	240
17	First Affiliated Hospital of Chongqing Medical University	41	41	GE Signa 3T	2000	40.0	240
19	Anhui Medical University	31	18	GE Signa 3T	2000	22.5	240
20	Southwest University	229	250	Siemens Tim Trio 3T	2000	30.0	242
21	Beijing Anding Hospital, Capital Medical University	65	79	Siemens Tim Trio 3T	2000	30.0	240
22	Second Xiangya Hospital, Central South University	20	18	Philips Gyroscan Achieva 3.0T	2000	30.0	250
23	West China Hospital, Sichuan University	23	22	Philips Achieva 3.0T TX	2000	30.0	240

Across the selected sites in the REST-meta-MDD consortium, the number of retained time points after preprocessing slightly varies due to site-specific acquisition protocols. While most sites provided 230 time points, others contributed data with 232 or 240 time points. These variations are a result of both initial protocol settings and preprocessing procedures (e.g., discarding initial scans to ensure magnetization equilibrium). A summary of time point lengths by site is given in Table A9.

Table A9: Post-preprocessing time point lengths across selected sites.

Site ID	Institution	Timepoints (after preprocessing)
15	Zhongda Hospital, Southeast University	230
17	First Affiliated Hospital of Chongqing Medical University	230
19	Anhui Medical University	230
20	Southwest University	232
21	Beijing Anding Hospital, Capital Medical University	230
22	Second Xiangya Hospital, Central South University	240
23	West China Hospital, Sichuan University	230

To ensure consistency in downstream dynamic sampling, all time series were uniformly truncated to the first 210 time points. This guarantees a consistent sampling space across all subjects regardless of site-specific acquisition length.

We then applied a sliding window with a fixed length of 70 TRs and a step size of 35 TRs. This configuration yields exactly five overlapping segments per subject, each representing a snapshot of short-term functional dynamics. These segments serve as candidate samples for evaluating structural stability and selecting representative subjects in our core-set selection framework.

We used 33 regions of interest (ROIs) associated with the default mode network (DMN), selected from the Dosenbach-160 atlas (Dosenbach et al., 2010) following the specification provided by Yan et al. (2019). These ROIs were identified using public scripts available at `https://github.com/Chaogan-Yan/PaperScripts/tree/master/Yan_2019_PNAS/Dos160` and used consistently in REST-meta-MDD-related studies.

M.2 Baseline Introduction and Parameter Settings

Baseline Introduction

- **Random**: Uniformly samples instances without considering model behavior or data statistics.
- **k-Means** (Hartigan & Wong, 1979): An unsupervised clustering algorithm that partitions data into k clusters based on feature similarity, with the core-set formed by selecting samples closest to the cluster centroids.

Table A10: List of 33 DMN-related ROIs selected from the Dosenbach-160 atlas.

ROI Index	ROI Type	Yeo Network (Label)
1	vmPFC	7 (DMN)
4	mPFC	7 (DMN)
5	aPFC	7 (DMN)
6	vmPFC	7 (DMN)
7	vmPFC	7 (DMN)
11	vmPFC	7 (DMN)
13	vmPFC	7 (DMN)
14	ACC	7 (DMN)
15	vlPFC	7 (DMN)
17	sup frontal	7 (DMN)
20	sup frontal	7 (DMN)
25	vFC	7 (DMN)
63	inf temporal	7 (DMN)
72	inf temporal	7 (DMN)
73	post cingulate	7 (DMN)
85	precuneus	7 (DMN)
90	post cingulate	7 (DMN)
91	inf temporal	7 (DMN)
93	post cingulate	7 (DMN)
94	precuneus	7 (DMN)
100	sup temporal	7 (DMN)
102	angular gyrus	7 (DMN)
104	IPL	7 (DMN)
105	precuneus	7 (DMN)
108	post cingulate	7 (DMN)
111	post cingulate	7 (DMN)
112	precuneus	7 (DMN)
115	post cingulate	7 (DMN)
117	angular gyrus	7 (DMN)
124	angular gyrus	7 (DMN)
132	precuneus	7 (DMN)
134	IPS	7 (DMN)
137	occipital	7 (DMN)

- **Forgetting** (Toneva et al., 2018): Ranks samples by the number of times they transition from correct to incorrect predictions during training.
- **Entropy** (Coleman et al., 2020): Scores samples using the entropy of model output probabilities to reflect prediction uncertainty.
- **Area Under the Margin (AUM)** (Pleiss et al., 2020): Computes the average margin between the true class probability and the highest non-true probability across epochs.
- **Example-Level L2 Norm (EL2N)** (Paul et al., 2021): Measures the L2 norm between model predictions and true labels over early training epochs as a proxy for example difficulty.
- **Coverage-Centric Selection (CCS)** (Zheng et al., 2022): Stratifies samples by importance scores (e.g., AUM) and performs balanced sampling across strata to preserve distributional coverage.
- **Evolution-aware Variance (EVA)** (Hong et al., 2024b): Aggregates prediction error variances within early and late training windows to capture evolving sample dynamics.
- **Balanced One-shot Subset Selection (BOSS)** (Acharya et al., 2024): Greedily selects a subset by maximizing a Beta-weighted objective over feature similarity, label variability, and difficulty-based scores.

Parameter Settings All baseline methods are evaluated using a unified training setup. The classification model is a compact residual network designed for multivariate time series inputs. It consists of a stem convolution followed by three residual blocks with output channels of 32, 64, and 128, respectively. A global average pooling layer and a fully connected classifier complete the architecture. This design balances expressiveness and efficiency for medium-scale time series classification.

The model is trained for 200 epochs using the Adam optimizer with a learning rate of 0.01, weight decay of 1e-4, and a batch size of 256. All methods, including full-data training and subset-based training, use identical configurations.

Per-sample importance scores are derived from training dynamics and used to evaluate a range of sampling strategies. For EVA, we compute the variance of the prediction error vector within two non-overlapping windows: epochs 100–109 and 190–199, following the original protocol.

Hyper parameters of **SCLCS** are optimized using grid searching in the flowing spaces:

1. Dimension of Transformer: [4, 8, 16, 32, 64, 128, 256].
2. Head Number: [2, 4, 8, 16, 32, 64].
3. Epochs for calculating **SPS**: [50, 100, 150, 200].

M.3 Environment

Experiments are performed on an 8-GPU (H20) high-performance computing cluster provided by the Large-scale Instrument Sharing Platform of Southwest University.

N Limitations

While this work establishes a new paradigm for efficient FC benchmarking, we identify several exciting avenues for future investigation that build upon our findings:

- **Deepening the SPS-SPI Connection:** Our results demonstrate a strong empirical link between low **SPS** (structural stability) and effective core-set selection. However, the precise theoretical mechanism connecting the training dynamics of our encoder to the performance ranking of diverse, external SPI models warrants deeper investigation. Future work could explore this link to build a more formal bridge between learnable latent structures and statistical model behavior.
- **Developing an Adaptive Sampling Strategy:** Our experiments show that the optimal choice between simple low-SPS ranking (**SCLCS**) and density-aware sampling (**SCLCS_Dense**) is task-dependent. A key next step is to develop heuristics or a meta-learning framework to automatically select the optimal sampling strategy based on dataset characteristics (e.g., class structure, sample heterogeneity), removing the need for manual selection.
- **Broader Generalization and Application:** While our evaluation on the large-scale, heterogeneous REST-meta-MDD dataset provides a strong test of robustness, the framework's generalizability should be further validated. A full study of supervision (site-ID, multi-task, and more) and how SPI properties modulate task difficulty is a valuable future direction. The relationship between the RP problem and data properties is another major extension to our work. Future studies should also apply **SCLCS** to fMRI datasets from different clinical populations (e.g., Alzheimer's disease, ADHD), other imaging modalities, or even entirely different domains of multivariate time-series analysis to establish the full scope of its applicability.
- **More Heterogeneous Scenarios:** Simply pulling resting-state and task-fMRI scans together into the contrastive loss would be problematic, as the functional state changes. This poses a more complex heterogeneous data fusion challenge. Exploring model benchmarking problem on heterogeneous scan types is a valuable future direction.

THE USE OF LARGE LANGUAGE MODELS (LLMS)

In the preparation of this manuscript, we utilize Large Language Models (LLMs) in a supportive capacity. Specifically, their use is confined to the following areas:

- **Writing and Editing:** LLMs are employed to enhance the clarity, grammar, and style of the text, ensuring the manuscript's readability.
- **Assistance with Theorem Proofs:** LLMs serves as an assistive tool to verify the logical consistency and correctness of individual steps within our mathematical derivations and proofs.

The core scientific ideas, the structure of the proofs, the experimental design, and all final conclusions presented in this paper are conceived and developed entirely by the authors.