# Dimensionality and Topological Stability of Neural Representations Predict Learning Outcomes

**Junjie Yu**[1] * **Zihan Deng**[1] * **Wenxiao Ma**[1] **Zhuoli Ouyang**[2] **Jianyu Zhang**[1]
**Yi Guo**[3] **Quanying Liu**[1] †
[1]Department of Biomedical Engineering, Southern University of Science and Technology
[2]Department of Electronic and Electrical Engineering, Southern University of Science and Technology
[3]Shenzhen People's Hospital, affiliated to Southern University of Science and Technology
{12231192, 12211804}@mail.sustech.edu.cn
xuanyi_guo@163.com
liuqy@sustech.edu.cn

## Abstract

Recent advances in deep learning suggest that the dimensionality of neural representations shapes generalization. We ask whether neural activity exhibits a similar principle during human learning. Using longitudinal fMRI collected from a real university course, we quantify representational geometry with intrinsic dimensionality, persistent homology, and Wasserstein distance. We find that learning outcomes depend on brain region: in association hubs supporting conceptual abstraction (e.g., Angular Gyrus), lower dimensionality predicts better performance, while in regions supporting complex perceptual processing (e.g., Temporal Fusiform Cortex), higher dimensionality predicts better performance. Topological analysis further shows that high-performing individuals form more stable structures in association hubs and that their representational topologies diverge more strongly from one another. Together, these findings suggest that effective learning in the brain relies on region-specific representational organization, with stable and individualized structures that support successful performance.

## 1 Introduction

The geometry of neural representations is closely linked to learning and generalization. In artificial neural networks (ANNs), a robust observation is that performance improves when information is encoded in lower-dimensional representations [1–5]. Lower-dimensional representations capture task-relevant structure while suppressing irrelevant variability, thereby facilitating generalization across diverse inputs.

The human brain, like ANNs, transforms external information into internal representations that support a wide range of tasks [6–9]. This parallel motivates a central question: **does the dimensionality of neural representations in the brain also predict generalization performance in learning** (Figure 1A)? To investigate this, we analyze a longitudinal fMRI dataset acquired while 20 students attended five weeks of online lectures, followed by a review week and a final exam [10]. fMRI was recorded throughout the entire course and exam, and exam scores provide an objective measure of learning outcomes. This design makes it possible to directly link representational dimensionality of course-related knowledge to behavioral performance.

*J. Yu and Z. Deng contributed equally.
†Corresponding author.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: AI for Science.

Our approach differs from prior work by focusing not on raw BOLD signals, but on their first-order temporal differences. This differential fMRI emphasizes dynamic changes in neural activity while mitigating the effects of slow fluctuations and background noise. The method highlights transient state transitions that are more directly linked to processes of information and knowledge encoding.

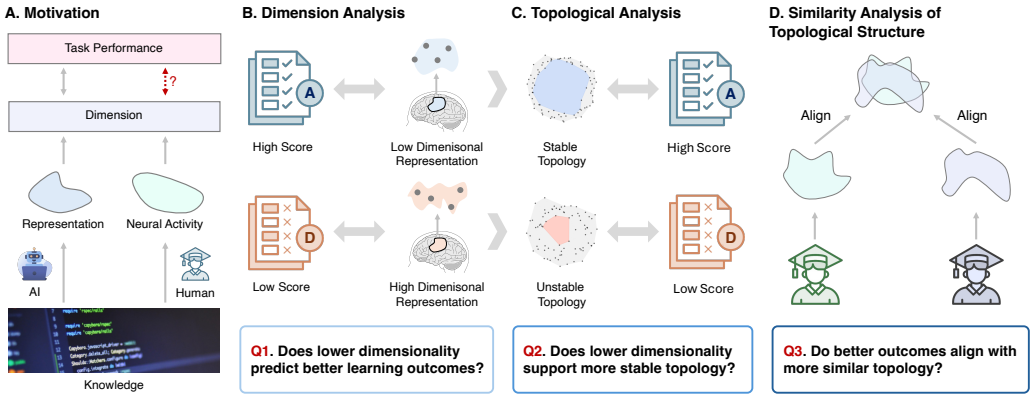


Figure 1: **Motivation and study framework. (A)** Motivation. Findings in artificial networks link representational dimensionality to generalization. We ask whether the dimensionality of human neural representations relates to learning outcomes, using exam scores as an objective index. **(B)** Dimension analysis. *Q1:* Does lower dimensionality predict better learning outcomes across cortical regions? **(C)** Topological analysis. *Q2:* In regions where lower dimensionality relates to better outcomes, do representations show more stable topology, quantified by longer $H_1$ lifetimes? **(D)** Similarity analysis. *Q3:* Among students with better outcomes, is representational topology more similar or more divergent across individuals, assessed by the 1-Wasserstein distance between persistence diagrams?

Using this framework, we analyze the relationship between neural representations and learning outcomes in three steps: first testing whether dimensionality predicts exam performance across regions (Figure 1B), then examining whether lower-dimensional regions form more stable topological structures (Figure 1C), and finally asking whether high-performing students converge to similar or diverge to more individualized topologies (Figure 1D).

This three-part analysis allows us to uncover systematic links between dimensionality, topology and learning outcome. Specifically, we find that the relationship between representational dimensionality and exam performance depends on brain region. In regions such as the temporal fusiform cortex, higher dimensionality correlates with better performance. In contrast, in the angular gyrus, lower dimensionality predicts stronger outcomes. Moreover, in regions showing negative correlations, representations exhibit stable topological structures, and high-performing students display greater inter-individual divergence in topology. These findings suggest that successful learners develop individualized representational geometries rather than converging on a uniform code.

In summary, the contributions of this work are:

- Introduction of differential fMRI as a method that enhances sensitivity to learning-related representational dynamics.
- Identification of region-specific associations between representational dimensionality and objective learning outcomes.
- Demonstration that in regions where lower dimensionality predicts stronger performance, representational topology becomes both stable and individualized in high-performing learners.

## 2 Related Work

**Dimensionality and generalization in deep learning.** Two complementary views link representation/parameter dimensionality to generalization. (i) *Embedding view.* Many studies show that

networks that compress internal embeddings generalize better, and that lower intrinsic dimensionality of representations relates to robustness and sample efficiency [1, 3–5]. (ii) *Parameter/trajectory view.* Independent lines of work measure the effective dimensionality of the optimization path or the low-loss subspaces of the objective, showing that gradient descent largely evolves within a tiny, low-dimensional subspace and that solutions lie in low-dimensional regions of the loss landscape [11–13]. Moreover, the lower the dimensionality of parameter-update trajectories, the stronger the models generalization ability [2].

**Low-dimensional neural representations in the brain.** Population activity in biological circuits often concentrates on low-dimensional manifolds that capture task-relevant latent variables and dynamics [14–16]. Moreover, low-dimensional organization has been linked to specific cognitive functions: prefrontal dynamics implementing context-dependent decision-making on a low-dimensional manifold [17], working-memory dynamics supported by low-dimensional attractors [18], and flexible sensorimotor computations via rapid reconfiguration of population trajectories [19]. **However, to our knowledge, no prior work has directly tested whether the dimensionality of human neural activity predicts individual learning outcomes.**

## 3 Methods

We analyze the geometry of neural representations during learning along three complementary dimensions:

1. Estimating the **intrinsic dimensionality** of neural activity across cortical regions and testing its relationship with learning outcomes.
2. Using **persistent homology** to quantify the stability of representational topology, and assessing whether lower-dimensional regions also exhibit more stable structures.
3. Measuring inter-individual similarity of topological structures with **Wasserstein distance** to evaluate whether high-performing learners converge or diverge in representational geometry.

All analyses are performed on *differential fMRI signals*, obtained by computing first-order temporal differences of raw BOLD time series. This preprocessing emphasizes rapid, learning-related state transitions while reducing the influence of slow fluctuations and background noise.

### 3.1 Intrinsic dimensionality estimation

The first analysis addresses how many degrees of freedom are required to describe neural activity during learning. The intrinsic dimensionality (ID) of a dataset reflects the effective number of latent variables that govern its structure. Low ID corresponds to compressed, constrained dynamics, while high ID indicates richer variability.

We estimate ID from differential fMRI time series $\{\Delta x_t\}$ using the maximum likelihood estimator (MLE) of Levina and Bickel [20]. For a point $x$, let $T_j(x)$ denote the Euclidean distance to its $j$-th nearest neighbor. The local estimator is

$$\hat{m}_k(x) = \left[ \frac{1}{k-1} \sum_{j=1}^{k-1} \log \left( \frac{T_k(x)}{T_j(x)} \right) \right]^{-1}. \tag{1}$$

Averaging over all samples yields the regional ID:

$$\bar{m}_k = \frac{1}{N} \sum_{i=1}^{N} \hat{m}_k(x_i). \tag{2}$$

This measure allows us to test whether compressed or expansive neural representations in specific brain regions are more predictive of learning outcomes.

### 3.2 Topological stability via persistent homology

Dimensionality captures compression, but not how representations are organized. To assess structural stability, we use persistent homology (PH), which tracks the appearance and disappearance of topological features (connected components, loops, voids) across scales.

Given a point cloud of neural states $\{\Delta x_t\}$, a filtration is built by gradually increasing a distance threshold $\epsilon$ and adding simplices when all edges are shorter than $\epsilon$. Persistent homology records when a feature (e.g., a loop in $H_1$) is *born* and when it *dies*. The difference $d - b$ defines its *lifetime*. **Long-lived features reflect stable organization, while short-lived ones are interpreted as noise.**

For each brain region, we compute the average lifetime of $H_1$ features as a summary of topological stability. This analysis tests whether lower-dimensional regions (from the previous step) also form more stable representational structures.

### 3.3 Inter-individual differences via Wasserstein distance

Finally, we ask whether successful learners converge to similar representations or instead diverge toward individualized geometries. For each participant, persistent homology produces a persistence diagram consisting of birth-death pairs $(b_i, d_i)$. This diagram is treated as an empirical distribution in $\mathbb{R}^2$, capturing both the scale and stability of topological features.

The distance between two participants diagrams $P$ and $Q$ is quantified using the 1-Wasserstein distance:

$$W_1(P, Q) = \inf_{\gamma \in \Gamma(P,Q)} \int_{\mathbb{R}^2 \times \mathbb{R}^2} \|u - v\| \, d\gamma(u, v), \tag{3}$$

where $\Gamma(P, Q)$ is the set of admissible couplings with marginals $P$ and $Q$.

Intuitively, this measures the minimal cost of transforming one topological structure into another. Small distances indicate convergence to similar representational structures, while large distances reflect divergence and individuality. This allows us to test whether high-performing learners develop common solutions or distinct representational geometries.

Together, these three analyses provide complementary perspectives on representational geometry during learning: intrinsic dimensionality characterizes compression, persistent homology reveals stability, and Wasserstein distance captures individuality.

### 3.4 Dataset and experimental design

We analyze data from the longitudinal fMRI study of Meshulam et al. [10], in which undergraduate students at Princeton University were scanned repeatedly while taking an introductory computer science course (COS 126). Over a 13-week semester, students underwent six fMRI sessions: five scans during lecture videos and one final scan with recap videos and an exam. Each lecture scan presented $\sim$40 minutes of video segments (3-5 per scan, total $\sim$197 minutes across the semester), while the final exam scan included five 3-minute recap videos followed by 16 open-ended exam questions.

The original study included 20 undergraduate students. All students had no prior background in computer science, and all were enrolled in the course for credit. Scanning was conducted on 3T Siemens MRI systems (Skyra/Prisma) with whole-brain coverage (3 mm isotropic voxels, TR = 2000 ms). In the present work, we focus specifically on the Week 2 scan, corresponding to the early stage of the semester.

## 4 Experiments and Results

### 4.1 Whole-Brain Dimensionality Distribution

We first computed intrinsic dimensionality across the cortex using differential fMRI signals (first-order temporal difference).

As shown in Figure 2, dimensionality varied systematically across cortex. Regions with the highest dimensionality included Heschl's Gyrus, Insular Cortex, Central Opercular Cortex and Frontal Oper-

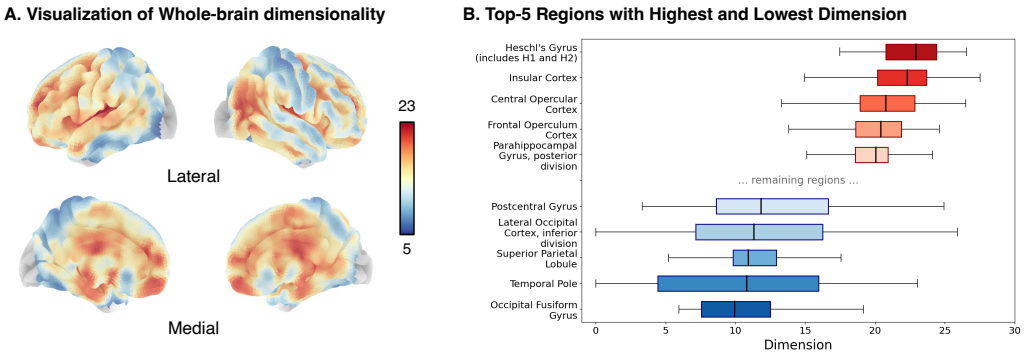


Figure 2: **Whole-brain intrinsic dimensionality.** **(A)** Cortical surface maps of dimensionality estimated from differential fMRI. **(B)** Dimensionality varies substantially across regions: highest in Heschls Gyrus and Insular Cortex, reflecting rich perceptual integration, and lowest in Occipital Fusiform Gyrus and Temporal Pole, reflecting compact, abstracted representations.

culum. These areas support auditory processing, multisensory integration and speech-motor control, where rich variability in neural codes may be required to represent continuous inputs.

By contrast, regions with the lowest dimensionality included the Occipital Fusiform Gyrus, Temporal Pole and Superior Parietal Lobule. These association areas are implicated in higher-level visual categorization, semantic integration, and spatial attention, functions that are consistent with more compact and abstracted codes.

### 4.2 Dimensionality of Differential fMRI Predicts Learning Outcomes

We estimated regional intrinsic dimensionality from differential fMRI and tested its relationship with exam performance (Fig. 3 A-B). Clear and region-specific associations emerged, with both positive and negative directions across cortex.

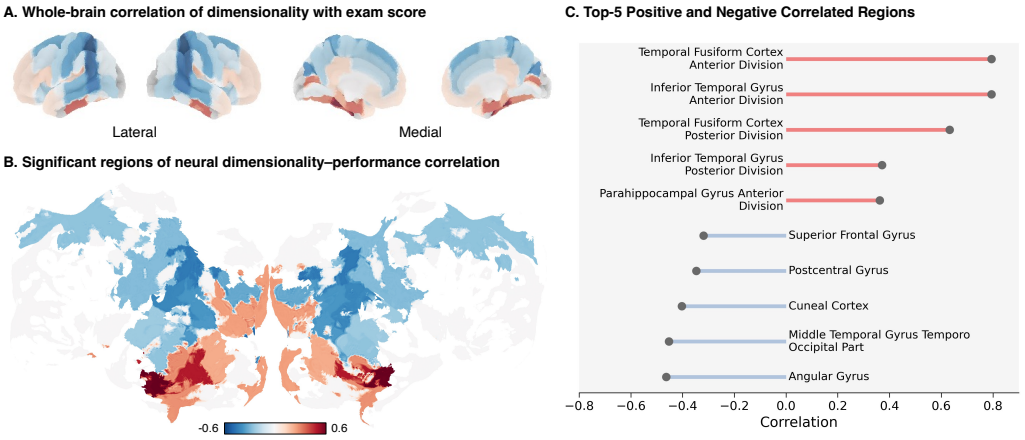


Figure 3: **Dimensionality-performance associations.** **(A)** Whole-brain maps of correlations between regional dimensionality and exam scores. **(B)** Regions showing significant correlations with exam performance. **(C)** The strongest effects show a bidirectional pattern: higher dimensionality benefits perceptual-temporal cortices (e.g., Temporal Fusiform), whereas lower dimensionality benefits association hubs (e.g., Angular Gyrus).

Perceptual and temporal regions, including the Temporal Fusiform Cortex (anterior and posterior divisions) and Inferior Temporal Gyrus showed positive correlations: higher dimensionality predicted better exam performance (Fig. 3C). Elevated dimensionality in these regions may provide richer and more flexible representational spaces that facilitate fine-grained visual categorization and semantic encoding.

By comparison, several association cortices exhibited negative correlations. The Angular Gyrus showed the strongest effect ($r \approx -0.46$), with a similar pattern observed in the Middle Temporal Gyrus. These areas are known to support semantic abstraction and multimodal integration. In this context, lower dimensionality may reflect compressed representational codes that emphasize task-relevant information while suppressing redundancy, thereby supporting efficient retrieval and generalization.

Together, these results indicate that dimensionality-performance relationships are heterogeneous across the brain. Perceptual and temporal cortices benefit from elaborated high-dimensional representations that capture the richness of sensory input, whereas association hubs benefit from compact low-dimensional codes that promote abstraction and integration. This raises the question of what structural property underlies the benefit of reduced dimensionality in association hubs.

### 4.3 Topological Stability Predicts Learning Outcomes

The dimensionality analyses above showed that in several association hubs, lower-dimensional representations predict better learning outcomes. A natural interpretation is that reduced dimensionality may reflect the emergence of more structured and organized representations, which in turn support abstraction and retrieval of knowledge. This raises two key questions: **(i) does the degree of structural organization itself predict learning outcomes**, and **(ii) do high-performing students converge to similar representational structures or instead form individualized ones?** Dimensionality alone cannot directly address these questions, as it quantifies compression but not the stability or organization of representational geometry.

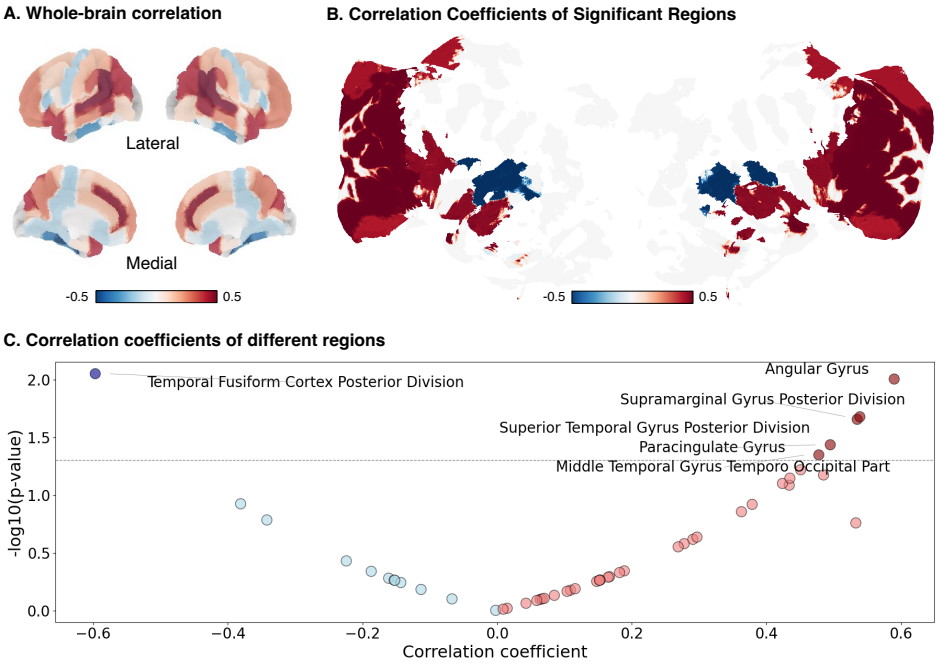


Figure 4: **Topological stability predicts performance.** **(A)** Whole-brain correlations between average $H_1$ lifetime and exam performance. **(B)** Significant regions show positive relations in association cortices (e.g., Angular, Supramarginal, Superior Temporal) and negative relations in Fusiform regions. **(C)** Correlation coefficients versus significance summarize that regions where more stable representational topology supports learning are also those where lower dimensionality was linked to performance.

To move beyond dimensionality, we turned to persistent homology, which captures the stability of representational topology by tracking the lifetime of loop-like ($H_1$) features across scales. We first tested whether topological stability predicts learning outcomes. Whole-brain analyses revealed widespread correlations between average $H_1$ lifetime and exam performance (Figure 4A–B). Association cortices including the Angular Gyrus, Supramarginal Gyrus, and Superior Temporal Gyrus

showed strong positive effects, indicating that successful learning in these hubs is supported by the formation of stable representational structures. In contrast, the Temporal Fusiform Cortex exhibited a significant negative correlation, consistent with its role in perceptual and semantic integration, where flexible but less stable structures may be advantageous.

Taken together, these results demonstrate that association hubs benefit from stable representational geometries that scaffold conceptual learning, whereas perceptual–temporal regions rely on richer but more variable structures. We next asked whether such stable structures converge across students or instead remain individualized.

### 4.4 Topological Divergence Across High-performing Students

To test whether stable structures converge or diverge across individuals, particularly among students with better learning outcomes, we compared persistence diagrams across participants using the 1-Wasserstein distance $W_1$, which quantifies the minimal cost of aligning two birth–death distributions of topological features. We then correlated pairwise $W_1$ values with the sum of exam scores of the corresponding student pair, such that higher values indicate that both students achieved better learning outcomes.

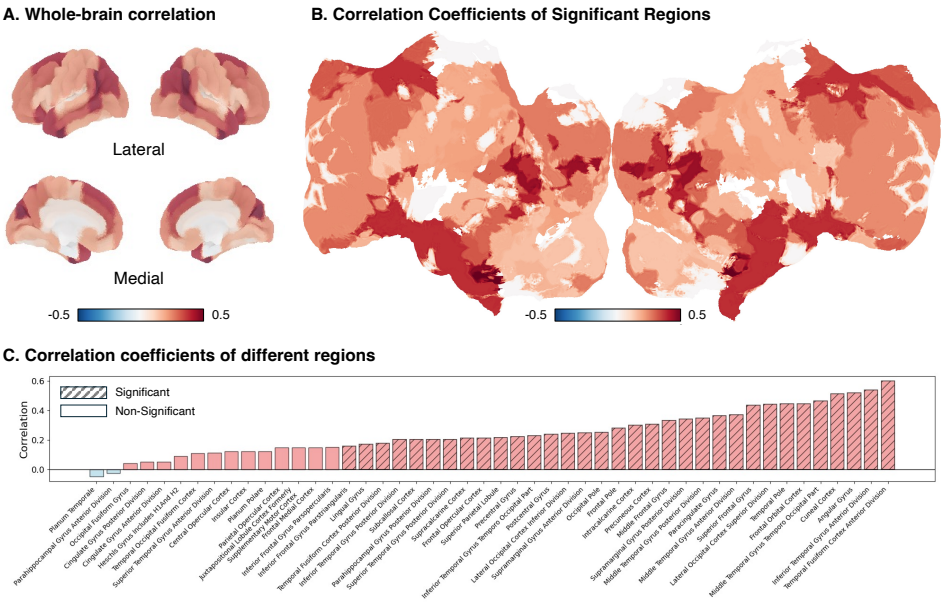


Figure 5: **Divergence of representational topology across individuals.** **(A)** Whole-brain correlations between pairwise 1-Wasserstein distance of persistence diagrams and the sum of exam scores. **(B)** Significant regions show that higher-performing pairs display larger divergence, reflecting more individualized representational organization. **(C)** Correlation coefficients versus significance demonstrate that effective learning produces neural codes that are both stable within individuals and heterogeneous across individuals.

As shown in Figure 5A–C, pairwise $W_1$ was positively correlated with sum of exam scores in most cortical regions. In other words, pairs of high-performing students exhibited more divergent topological structures than pairs of lower-performing students.

Taken together, stability within individuals and divergence across individuals emerge as complementary principles of efficient brain organization during learning.

## 5 Discussion and Conclusion

Our results show that the relationship between representational geometry and learning is systematically organized across cortex. Perceptual and temporal regions benefited from higher dimensionality, consistent with the need to encode rich and fine-grained variability in sensory streams. In contrast,

association hubs, most prominently the Angular Gyrus, benefited from lower dimensionality and exhibited more stable topological structure, suggesting compressed and reliable representations of conceptual knowledge. These effects were observed across complementary analyses of intrinsic dimensionality, persistent homology, and inter-individual topology, with the Angular Gyrus consistently emerging as one of the strongest loci.

The repeated involvement of the Angular Gyrus is unlikely to be incidental. This region has been identified as a multimodal integration hub within the default mode and semantic networks, supporting the binding of distributed features into coherent concepts and contributing to episodic recollection and attention to memory [21–23]. These established roles help explain why our three empirical findings converge on this region. First, lower dimensionality in the Angular Gyrus predicted better performance, consistent with its function as a convergence zone that compresses diverse inputs into abstract and behaviorally useful codes [24]. Second, greater topological stability in the Angular Gyrus was associated with more successful learning, suggesting that effective learners form persistent representational structures that provide a scaffold for retrieval and generalization [25]. Third, high-performing students exhibited within-individual stability but across-individual divergence in representational organization, a pattern that is consistent with the Angular Gyrus supporting individualized conceptual schemas that are stable for each learner yet need not align across people [26, 27].

These observations refine comparisons between biological and artificial systems. Artificial neural networks often achieve better generalization through compressed embeddings. In the brain, however, expansion is advantageous in sensory regions where preserving input richness is essential, while compression is advantageous in association hubs such as the Angular Gyrus where abstraction and integration are required. Successful learning therefore reflects a flexible balance between expansion and compression across cortical regions.

Several limitations should be acknowledged. Our analyses are correlational, were conducted within a single course context, and involved a modest sample size. Future work should examine longitudinal changes across the full timescale of learning, test whether subdivisions of the Angular Gyrus contribute differentially, and investigate causal mechanisms using perturbation or stimulation approaches. Extending the framework to broader populations and diverse tasks will further clarify the generality of the principles identified here.

In conclusion, we provide convergent evidence that representational dimensionality and topological stability jointly predict human learning outcomes. Effective learning involves region-specific optimization, with the Angular Gyrus emerging as a keystone for integrating and compressing conceptual knowledge. High-performing learners form neural representations that are topologically stable within individuals yet divergent across individuals, linking individualized learning strategies with stable representational scaffolds. These findings bridge biological and artificial systems and highlight representational geometry as a foundation for general learning.

## References


[1] Stefano Recanatesi, Matthew Farrell, Madhu Advani, Timothy Moore, Guillaume Lajoie, and Eric Shea-Brown. Dimensionality compression and expansion in deep neural networks. *arXiv preprint arXiv:1906.00443*, 2019.

[2] Umut Simsekli, Ozan Sener, George Deligiannidis, and Murat A Erdogdu. Hausdorff dimension, heavy tails, and generalization in neural networks. *Advances in Neural Information Processing Systems*, 33:5138–5151, 2020.

[3] Alessio Ansuini, Alessandro Laio, Jakob H Macke, and Davide Zoccolan. Intrinsic dimension of data representations in deep neural networks. *Advances in Neural Information Processing Systems*, 32, 2019.

[4] Tolga Birdal, Aaron Lou, Leonidas J Guibas, and Umut Simsekli. Intrinsic dimension, persistent homology and generalization in neural networks. *Advances in neural information processing systems*, 34:6776–6789, 2021.

[5] Armen Aghajanyan, Luke Zettlemoyer, and Sonal Gupta. Intrinsic dimensionality explains the effectiveness of language model fine-tuning. *arXiv preprint arXiv:2012.13255*, 2020.

[6] Alex Martin. The representation of object concepts in the brain. *Annu. Rev. Psychol.*, 58(1): 25–45, 2007.

[7] Anja Ischebeck, Michael Schocke, and Margarete Delazer. The processing and representation of fractions within the brain: An fmri investigation. *NeuroImage*, 47(1):403–413, 2009.

[8] Marie L Smith, Frédéric Gosselin, and Philippe G Schyns. Measuring internal representations from behavioral and brain data. *Current Biology*, 22(3):191–196, 2012.

[9] Andrew C Connolly, J Swaroop Guntupalli, Jason Gors, Michael Hanke, Yaroslav O Halchenko, Yu-Chien Wu, Hervé Abdi, and James V Haxby. The representation of biological classes in the human brain. *Journal of Neuroscience*, 32(8):2608–2618, 2012.

[10] Meir Meshulam, Liat Hasenfratz, Hanna Hillman, Yun-Fei Liu, Mai Nguyen, Kenneth A Norman, and Uri Hasson. Neural alignment predicts learning outcomes in students taking an introduction to computer science course. *Nature communications*, 12(1):1922, 2021.

[11] Chunyuan Li, Heerad Farkhoor, Rosanne Liu, and Jason Yosinski. Measuring the intrinsic dimension of objective landscapes. *arXiv preprint arXiv:1804.08838*, 2018.

[12] Guy Gur-Ari, Daniel A Roberts, and Ethan Dyer. Gradient descent happens in a tiny subspace. *arXiv preprint arXiv:1812.04754*, 2018.

[13] Jonathan Frankle, Gintare Karolina Dziugaite, Daniel Roy, and Michael Carbin. Linear mode connectivity and the lottery ticket hypothesis. In *International Conference on Machine Learning*, pages 3259–3269. PMLR, 2020.

[14] Mark M Churchland, John P Cunningham, Matthew T Kaufman, Justin D Foster, Paul Nuyujukian, Stephen I Ryu, and Krishna V Shenoy. Neural population dynamics during reaching. *Nature*, 487(7405):51–56, 2012.

[15] Juan A Gallego, Matthew G Perich, Lee E Miller, and Sara A Solla. Neural manifolds for the control of movement. *Neuron*, 94(5):978–984, 2017.

[16] John P Cunningham and Byron M Yu. Dimensionality reduction for large-scale neural recordings. *Nature neuroscience*, 17(11):1500–1509, 2014.

[17] Valerio Mante, David Sussillo, Krishna V Shenoy, and William T Newsome. Context-dependent computation by recurrent dynamics in prefrontal cortex. *nature*, 503(7474):78–84, 2013.

[18] Mark S Goldman. Memory without feedback in a neural network. *Neuron*, 61(4):621–634, 2009.

[19] Evan D Remington, Devika Narain, Eghbal A Hosseini, and Mehrdad Jazayeri. Flexible sensorimotor computations through rapid reconfiguration of cortical dynamics. *Neuron*, 98(5): 1005–1019, 2018.

[20] Elizaveta Levina and Peter Bickel. Maximum likelihood estimation of intrinsic dimension. *Advances in neural information processing systems*, 17, 2004.

[21] Mohamed L Seghier. The angular gyrus: multiple functions and multiple subdivisions. *The Neuroscientist*, 19(1):43–61, 2013.

[22] Jeffrey R Binder and Rutvik H Desai. The neurobiology of semantic memory. *Trends in cognitive sciences*, 15(11):527–536, 2011.

[23] Michael D Rugg and Kaia L Vilberg. Brain networks underlying episodic memory retrieval. *Current opinion in neurobiology*, 23(2):255–260, 2013.

[24] Amy R Price, Michael F Bonner, Jonathan E Peelle, and Murray Grossman. Converging evidence for the neuroanatomic basis of combinatorial semantics in the angular gyrus. *Journal of Neuroscience*, 35(7):3276–3284, 2015.

[25] Heidi M Bonnici, Franziska R Richter, Yasemin Yazar, and Jon S Simons. Multimodal feature integration in the angular gyrus during episodic and semantic retrieval. *Journal of Neuroscience*, 36(20):5462–5471, 2016.

[26] Rodrigo M Braga and Randy L Buckner. Parallel interdigitated distributed networks within the individual estimated by intrinsic functional connectivity. *Neuron*, 95(2):457–471, 2017.

[27] Deniz Vatansever, David K Menon, and Emmanuel A Stamatakis. Default mode contributions to automated information processing. *Proceedings of the National Academy of Sciences*, 114 (48):12821–12826, 2017.