Neuro-Anatomy–Informed Self-Supervised Learning for Structural Brain MRI

Sean Chen Jiale [*1] **Tianze Yu** [*1] **Somayeh Ebrahimkhani** [*1] **Lin Zhi Zheng Shawn** [2 3] **Kok Pin Ng** [2 3] **Ngai-Man Cheung** [1]

Abstract

Self-supervised learning (SSL) has become a foundational paradigm for representation learning from large-scale unlabeled data and underpins many modern foundation models across biological and biomedical domains, including biomedical imaging. However, most existing SSL approaches for biomedical imaging rely on objectives originally designed for natural images, such as contrastive alignment or reconstruction, which primarily enforce instance-level consistency based on generic features. As a result, these objectives are insufficient to capture biologically meaningful structural relationships that emerge across samples and populations, limiting their ability to learn representations aligned with underlying biological organization. **In this work**, we introduce neuro-anatomy-informed self-supervised learning (NAI-SSL), a framework that moves beyond conventional instance-level objectives by incorporating biologically grounded inductive bias. Focusing on structural brain MRI, we leverage *inter-hemispheric structural consistency*, a fundamental principle of brain organization, to formulate a structure-aware self-supervised task based on inter-regional covariance. Unlike conventional SSL objectives that operate on individual samples, the proposed task leverages cross-sample covariance statistics to model coordinated variation across anatomically related brain regions, enabling the learning of *biologically meaningful cross-sample structural relationships*. Empirical results demonstrate that the proposed approach consistently improves representation quality across downstream neuroimaging tasks, while producing representations that better reflect underlying brain organization. **Additional analysis and experiments are included in the Supp.**

[*]Equal contribution [1]Singapore University of Technology and Design, Singapore [2]Department of Neurology, National Neuroscience Institute, Singapore [3]Duke-NUS Medical School, Singapore. Correspondence to: Ngai-Man Cheung (Project Lead) <ngaiman_cheung@sutd.edu.sg>, Kok Pin Ng <ng.kok.pin@singhealth.com.sg>.

Proceedings of the ICML 2026 3rd Workshop on Multi-modal Foundation Models and Large Language Models for Life Sciences, Seoul, Korea. 2026. Copyright 2026 by the author(s).

1. Introduction

Self-supervised learning (SSL) (Chen et al., 2020; He et al., 2022; Tong et al., 2022; Assran et al., 2023; Bardes et al., 2024) has emerged as a powerful paradigm for learning representations from large-scale unlabeled data and forms a key foundation of many modern foundation models. SSL has recently been adopted for biomedical imaging (Zhou et al., 2023b; Tak et al., 2026; Liu et al., 2026; Mazher et al., 2025; Caro et al.; Shakouri et al., 2023; Dong et al., 2024; Xu et al., 2025), where expert annotation is costly and often constrained by the need for specialized clinical expertise. Applications span a wide range of domains, including neuroimaging (Tak et al., 2026; Xu et al., 2025; Yang et al., 2025), retinal imaging (Liu et al., 2026; Zhou et al., 2023b; Yu et al., 2024) and computed tomography (CT) (Zhuang et al., 2025b; Li et al., 2025; Valanarasu et al., 2024), where SSL enables scalable representation learning from large collections of unlabeled data. Existing SSL approaches for biomedical imaging can be broadly grouped by their underlying learning principles such as contrastive learning, reconstruction, latent prediction and self distillation. Contrastive learning has been widely explored in biomedical imaging. For example, BrainIAC (Tak et al., 2026) adopts a SimCLR-style contrastive objective (Chen et al., 2020) to learn invariant representations across augmented brain MRI views. Reconstruction-based SSL has also shown strong effectiveness in this domain. RETFound (Zhou et al., 2023b), for instance, learns generalizable retinal representations via masked autoencoder (MAE) pretraining (He et al., 2022). See Sec. A for detailed related work.

Research Gap. Despite recent methodological advances, *most SSL approaches for biomedical imaging still rely on objectives originally developed for natural images* (See Sec. A for details). These approaches predominantly adopt general-purpose formulations with limited consideration of the underlying biological structure. More importantly, these objectives provide supervision primarily at the level of local appearance or instance-wise representations, without ex-

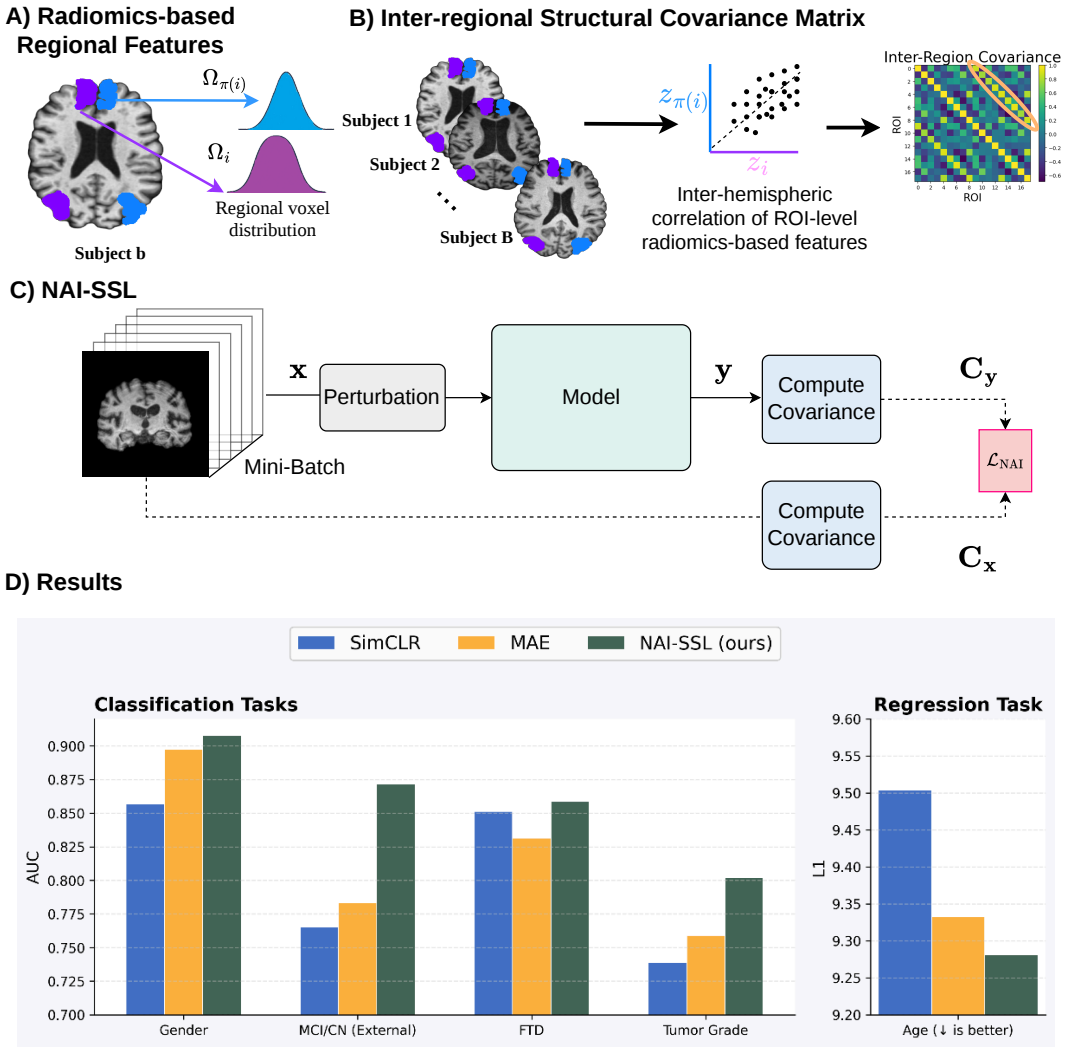

Figure 1. **Overview of the proposed NAI-SSL framework.** (A) For each subject, ROI-level radiomics features are extracted from anatomically defined brain regions. (B) Across a batch of subjects, inter-regional covariance is computed from ROI-level features, capturing coordinated anatomical variation across brain regions. The highlighted entries within the left–right covariance block correspond to homologous inter-hemispheric region pairs, which define the proposed covariance-based supervision. (C) During pretraining, input MRI volumes are perturbed and processed by the model to produce outputs from which inter-regional covariance statistics are computed. NAI-SSL enforces consistency between covariance structures computed from the input and model output, enabling the model to learn biologically meaningful cross-sample structural relationships beyond conventional instance-level objectives. (D) Downstream evaluation across representative tasks shows that NAI-SSL consistently improves representation quality.

plicitly modeling cross-sample inter-regional relationships governed by biological organization. This limitation is particularly critical in neuroimaging, where brain structure exhibits well-defined organization with meaningful dependencies across regions.

Our contributions. *Can self-supervised learning be improved by grounding pretext task design in scientific principles governing the data, beyond general-purpose objectives?*

We address this question in the context of structural brain MRI, a domain characterized by well-established neuroanatomical organization. Structural MRI is widely used for studying brain morphology and neurodegenerative diseases, enabling quantitative analysis of regional brain structure across large populations. Importantly, neuroscience has established that brain structure follows consistent organizational principles, among which *inter-hemispheric structural consistency* is one of the most prominent (Alexander-Bloch et al., 2013; Bullmore & Sporns, 2009), reflecting coordinated variation across anatomically related regions. We focus on this principle as a representative and well-established form of global brain organization for guiding representation learning.

In this paper, we propose **neuro-anatomy-informed self-supervised learning (NAI-SSL)**, *a framework that grounds the design of self-supervised tasks in neuroanatomical principles* (Fig. 1). In contrast to existing approaches that rely on instance-level objectives, we define a structure-aware self-supervised task based on inter-hemispheric structural consistency. This task replaces instance-level supervision with a structured statistical signal that captures *cross-sample coordinated variation across brain regions*, which cannot be inferred from individual samples alone, enabling the learned representations to better capture biologically meaningful structure. We conduct extensive experiments across a diverse set of structural brain MRI tasks and demonstrate that the proposed NAI-SSL consistently outperforms representative SSL approaches. Qualitative analyses based on attention visualization further show that the learned representations exhibit improved biological and clinical grounding.

2. Method

2.1. Neuroanatomical Principle: Inter-hemispheric Structural Consistency

The design of an effective self-supervised task critically depends on identifying intrinsic structure in the data that can provide meaningful supervision. In the context of structural brain MRI, extensive studies have established that brain organization follows well-defined neuroanatomical principles (Alexander-Bloch et al., 2013; Xu et al., 2018). Among these, inter-hemispheric structural consistency is one of the most prominent, reflecting coordinated variation between anatomically corresponding regions across the two hemispheres (Alexander-Bloch et al., 2013; Carmon et al., 2020).

This principle is closely related to the notion of structural covariance, whereby morphological properties of different brain regions exhibit correlated variation across individuals (Carmon et al., 2020; Chao et al., 2024; Mongay-Ochoa et al., 2025). Such covariance patterns have been widely observed in large-scale neuroimaging studies and are understood to reflect underlying developmental, genetic, and functional relationships (Alexander-Bloch et al., 2013; Gong et al., 2012; Mongay-Ochoa et al., 2025). Importantly, these relationships are inherently defined *across samples*, rather than within individual images.

Beyond its role as a fundamental organizational principle, inter-hemispheric structural consistency is also relevant in clinical settings. Neurological and neurodegenerative conditions are often associated with disruptions in coordinated structure between hemispheres, which may manifest as asymmetric patterns of structural change (Mongay-Ochoa et al., 2025; Brown et al., 2025; Li et al., 2023). Such deviations from typical inter-hemispheric relationships provide clinically meaningful signals that are not captured by local region analysis (Seeley, 2010; Vogel et al., 2023).

Motivated by these observations, we design our self-supervised task to capture inter-regional covariance patterns that encode inter-hemispheric structural consistency explicitly. By requiring the model to preserve these relationships, the task encourages the learning of representations that reflect biologically meaningful organization (Seeley, 2010; Vogel et al., 2023).

2.2. Self-Supervised Task Formulation

Region-level representation. Let $\mathbf{x} \in \mathbb{R}^{H\times W\times D}$ denote a 3D MRI volume, and let $\mathbf{y}$ denote the corresponding model output. We partition the brain into R anatomically defined ROIs using a neuroanatomical atlas (Fig. 1). Let Ω_i denote the spatial region corresponding to the i-th ROI. We extract a feature from each ROI, yielding $\mathbf{z} = [z_1, z_2, \ldots, z_R]^\top \in \mathbb{R}^R$, where z_i is computed based on Ω_i. The specific feature design is discussed in Supp. B.

Inter-regional covariance. Given a minibatch of B samples, we stack their region-level features into a data matrix $\mathbf{X} \in \mathbb{R}^{B\times R}$, where each row corresponds to a sample and each column corresponds to a region. We center and normalize $\mathbf{X}$ across the batch by applying zero-mean and unit-variance normalization to each column (region), yielding $\tilde{\mathbf{X}}$. We compute inter-regional covariance matrix as

$$\mathbf{C} = \frac{1}{B-1}\tilde{\mathbf{X}}^\top\tilde{\mathbf{X}} \in \mathbb{R}^{R\times R}, \tag{1}$$

which captures relationships between brain regions across samples.

Self-supervised task based on inter-hemispheric structural consistency. We partition the regions into left and right hemispheres, denoted by index sets $\mathcal{L}$ and $\mathcal{R}$, with $|\mathcal{L}| = |\mathcal{R}| = R/2$. Let $\pi : \mathcal{L} \to \mathcal{R}$ denote the mapping from each region in the left hemisphere to its homologous counterpart in the right hemisphere, as defined by the neuroanatomical atlas. These *homologous pairs* $(\Omega_i, \Omega_{\pi(i)})$ correspond to anatomically matched regions and are known to exhibit strong inter-hemispheric structural consistency (Alexander-Bloch et al., 2013; Xu et al., 2018). By focusing on these pairs, we capture biologically meaningful cross-region relationships while reducing spurious correlations introduced by unrelated region pairs.

Let $\mathbf{C}_\mathbf{x}$ and $\mathbf{C}_\mathbf{y}$ denote the covariance matrices computed from the input $\mathbf{x}$ and model output $\mathbf{y}$, respectively. We define the proposed self-supervised task as preserving inter-hemispheric covariance structure across samples over homologous regions. This is implemented by enforcing agreement in inter-hemispheric covariance between the input and model output:

$$\mathcal{L}_{\text{NAI}} = \sum_{i\in\mathcal{L}} \left(C_\mathbf{x}^{i,\pi(i)} - C_\mathbf{y}^{i,\pi(i)} \right)^2. \tag{2}$$

Table 1. Comparison of SSL pretext strategies evaluated on downstream and external test tasks. The best result per column is **bolded** and the second-best is underlined. BAcc = balanced accuracy (%).

	Age	Gender			MCI/CN			FTD/CN		
Pretext	L1↓	BAcc↑	AUC↑	F1↑	BAcc↑	AUC↑	F1↑	BAcc↑	AUC↑	F1↑
MAE	9.3324	81.77	0.8975	0.8199	71.65	0.8053	0.7254	78.33	0.8314	0.7735
SimCLR	9.5040	80.71	0.8569	0.8086	70.57	0.7724	0.7129	82.30	0.8511	0.8145
NAI-SSL	**9.2811**	**82.84**	**0.9076**	**0.8313**	**78.05**	**0.8261**	**0.7871**	**83.04**	**0.8585**	**0.8148**

	Tumor Grade			AD/MCI/CN (ext)			Age (ext)	MCI/CN (ext)		
Pretext	BAcc↑	AUC↑	F1↑	BAcc↑	AUC↑	F1↑	L1↓	BAcc↑	AUC↑	F1↑
MAE	61.79	0.7589	0.6330	48.35	0.7813	0.4732	9.6006	66.98	0.7833	0.6506
SimCLR	57.86	0.7389	0.5891	**56.65**	0.7537	0.5142	11.4789	62.02	0.7653	0.5893
NAI-SSL	**65.09**	**0.8021**	**0.6717**	55.00	**0.8004**	**0.5485**	**9.5657**	**78.91**	**0.8716**	**0.6984**

Table 2. We evaluate on eight downstream tasks covering neurodegenerative disease classification (MCI/CN, AD/MCI/CN, FTD/CN), brain-age regression, gender classification, tumor grading, and external evaluation benchmarks (denoted as ext). The best result in each column is highlighted in **bold** and the second-best is underlined. BAcc = balanced accuracy (%).

	Age	Gender			MCI/CN			FTD/CN		
Model	L1↓	BAcc↑	AUC↑	F1↑	BAcc↑	AUC↑	F1↑	BAcc↑	AUC↑	F1↑
BrainIAC (Tak et al., 2026)	10.700	51.82	0.597	0.424	59.70	0.745	0.578	79.11	**0.880**	0.778
BrainSegFounder (Cox et al., 2024)	15.270	77.52	0.858	0.775	62.54	0.695	0.619	73.72	0.793	0.732
MedicalNet (Chen et al., 2019)	15.370	78.89	0.888	0.793	67.12	0.744	0.609	62.78	0.692	0.626
DAE (Valanarasu et al., 2024)	13.403	80.46	0.894	0.807	71.44	0.799	0.727	75.14	0.867	0.755
MiM (Zhuang et al., 2025b)	15.089	75.25	0.889	0.753	50.00	0.818	0.389	59.98	0.766	0.579
3DINO-ViT (Xu et al., 2025)	9.853	**83.09**	0.900	**0.833**	66.93	0.786	0.677	79.62	0.865	0.793
Triad (Wang et al., 2026)	14.994	78.08	0.886	0.783	77.71	**0.835**	0.787	73.07	0.795	0.711
AMAES (Munk et al., 2024)	15.506	78.89	0.884	0.793	68.60	0.743	0.695	77.02	0.827	0.761
3D-Neuro-SimCLR (Kaczmarek et al., 2025)	13.891	57.45	0.801	0.457	67.33	0.761	0.681	59.75	0.833	0.496
NAI-SSL (Ours)	**9.281**	82.84	**0.908**	0.831	**78.05**	0.826	**0.787**	**83.04**	0.859	**0.815**

	Tumor Grade			AD/MCI/CN (ext)			Age (ext)	MCI/CN (ext)		
Model	BAcc↑	AUC↑	F1↑	BAcc↑	AUC↑	F1↑	L1↓	BAcc↑	AUC↑	F1↑
BrainIAC (Tak et al., 2026)	61.78	0.657	0.633	43.14	0.782	0.403	11.160	**82.71**	0.833	**0.731**
BrainSegFounder (Cox et al., 2024)	57.19	0.696	0.580	33.02	0.650	0.307	13.868	55.74	0.777	0.576
MedicalNet (Chen et al., 2019)	53.54	0.563	0.430	41.75	0.618	0.420	13.960	67.36	0.791	0.568
DAE (Valanarasu et al., 2024)	48.67	0.603	0.437	40.51	0.693	0.415	12.801	62.02	0.606	0.589
MiM (Zhuang et al., 2025b)	62.49	0.656	0.654	33.00	0.750	0.308	13.871	50.00	0.755	0.483
3DINO-ViT (Xu et al., 2025)	60.49	**0.819**	0.622	54.03	0.785	**0.553**	10.088	69.61	0.787	0.656
Triad (Wang et al., 2026)	61.47	0.720	0.492	34.61	0.583	0.334	13.710	49.77	0.499	0.481
AMAES (Munk et al., 2024)	50.00	0.534	0.444	43.20	0.726	0.403	14.045	49.77	0.753	0.482
3D-Neuro-SimCLR (Kaczmarek et al., 2025)	57.23	0.722	0.579	47.98	0.798	0.496	10.156	58.6	0.796	0.604
NAI-SSL (Ours)	**65.09**	0.802	**0.672**	**55.00**	**0.800**	0.549	**9.566**	78.91	**0.872**	0.698

Discussion. This formulation explicitly focuses on cross-hemisphere covariance between homologous regions, corresponding to selected off-diagonal entries (i.e., cross-hemisphere covariance between homologous regions) rather than diagonal terms. By restricting supervision to anatomically matched pairs, the proposed task provides a cleaner and more biologically grounded learning signal, avoiding spurious correlations between unrelated regions. **A thorough discussion of key design choices in the proposed task, including region-level feature, covariance estimation, and implementation are given in Supp. B.**

3. Experimental Evaluation

3.1. Experimental Settings

We pretrain NAI-SSL on a collection of structural brain MRI scans aggregated from multiple public datasets, totaling 12,541 MRIs. All scans are skull-stripped, bias-field corrected, and registered to Montreal Neurological Institute (MNI) (Collins, 2009) space. These unlabeled data are used exclusively for self-supervised pretraining. For downstream evaluation, we closely follow the BrainIAC (Tak et al., 2026) protocol. Representation quality is evaluated using the standard SSL linear probing protocol across **eight** downstream tasks under identical settings. See Supp. D for details.

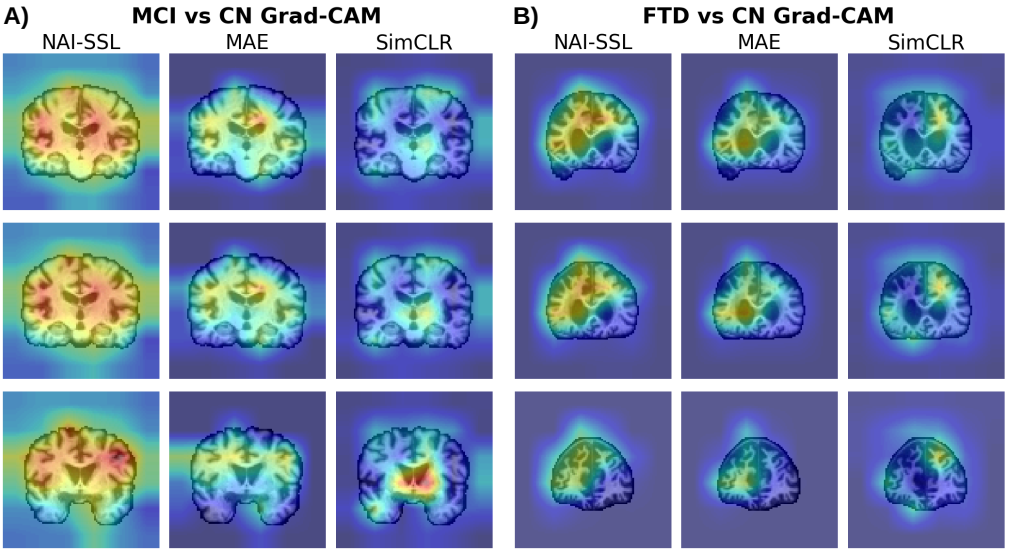

Figure 2. **Grad-CAM analysis of classification tasks. (A)** MCI vs. CN and **(B)** FTD vs. CN. NAI-SSL demonstrates more symmetric whole-brain attention and greater localization to neuroanatomical regions relevant to Alzheimer's disease (cingulate cortex, hippocampi, temporal cortex, parietal lobule, thalamus) and frontotemporal dementia (cingulate cortex, lateral and dorsomedial prefrontal cortex).

3.2. NAI-SSL outperforms standard SSL objectives and SOTA biomedical foundation models

We compare NAI-SSL against representative SSL pretext strategies (Tab. 1). As shown in Tab. 1, comparing to MAE and SimCLR, NAI-SSL improves on all tasks. We note that all SSL methods are pretrained for 50 epochs with identical details and hyperparameters, see Supp. C–E.

We further compare NAI-SSL against SOTA biomedical SSL methods and foundation models. From the results in Tab. 2, across the 20 reported metrics, NAI-SSL achieves the best result on 12 metrics and the second-best result on 7 more. These gains are especially pronounced on the clinically important neurodegenerative classification tasks, suggesting that the proposed pretext task produces more transferable disease-relevant representations than standard SSL objectives. We note that evaluation benchmarks overlap with pretraining datasets of BrainIAC (Tak et al., 2026) and 3DINO-ViT (Xu et al., 2025). Nevertheless, as shown in Tab. 2, NAI-SSL achieves the best or second-best performance on the majority of the evaluated tasks and outperforms prior biomedical SSL models, including BrainIAC (Tak et al., 2026), MiM (Zhuang et al., 2025b), and other strong SSL models. Overall, these results show that *grounding self-supervision in neuroanatomical structure consistently improves over standard SSL objectives.*

3.3. NAI-SSL produces more biologically meaningful attention patterns

Fig. 2 compares Grad-CAM activation maps across models pretrained with different SSL objectives. Relative to MAE and SimCLR, NAI-SSL produces more spatially coherent and anatomically consistent attention patterns across both hemispheres. In MCI/CN and FTD/CN tasks, NAI-SSL highlights broader, clinically relevant cortical and subcortical regions rather than isolated local areas. *Importantly, these maps were independently verified by a certified radiologist, confirming their biological interpretability and clinical relevance.*

4. Conclusion

In this work, we introduced NAI-SSL, a framework that grounds self-supervised task design in neuroanatomical principles for structural brain MRI. Unlike conventional SSL objectives that primarily operate at the level of individual samples or instance-wise representations, the proposed formulation models biologically meaningful cross-sample structural relationships through inter-regional covariance consistency between homologous brain regions. We analyzed design choices for NAI-SSL. Extensive experiments across diverse downstream tasks demonstrate that NAI-SSL consistently improves representation quality over representative SSL paradigms and achieves state-of-the-art performance among structural brain MRI representation learning methods. Qualitative analyses further show improved biological grounding in the learned representations. *Overall, this work provides evidence that grounding SSL task design in domain-specific scientific structure can improve representation learning in scientifically structured domains beyond general-purpose objectives.*

Acknowledgment

This research is supported by the National Research Foundation, Singapore under its AI Singapore Programmes (AISG Award No.: AISG2-TC-2022-007); The Agency for Science, Technology and Research (A*STAR) under its MTC Programmatic Funds (Grant No. M23L7b0021). This research is supported by the National Research Foundation, Singapore and Infocomm Media Development Authority under its Trust Tech Funding Initiative. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not reflect the views of National Research Foundation, Singapore and Infocomm Media Development Authority. The work is sponsored by the SUTD Decentralised Gap Funding Grant.

References

Alexander-Bloch, A., Giedd, J. N., and Bullmore, E. Imaging structural co-variance between human brain regions. *Nature reviews neuroscience*, 14(5):322–336, 2013.

Assran, M., Duval, Q., Misra, I., Bojanowski, P., Vincent, P., Rabbat, M., LeCun, Y., and Ballas, N. Self-supervised learning from images with a joint-embedding predictive architecture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15619–15629, June 2023.

Barbano, C. A., Brunello, M., Dufumier, B., and Grangetto, M. Anatomical foundation models for brain mris. *Pattern Recognition Letters*, 199:178–184, 2026. ISSN 0167-8655. doi: https://doi.org/10.1016/j.patrec.2025.11.028. URL https://www.sciencedirect.com/science/article/pii/S0167865525003848.

Bardes, A., Garrido, Q., Ponce, J., Chen, X., Rabbat, M., LeCun, Y., Assran, M., and Ballas, N. Revisiting feature prediction for learning visual representations from video. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=QaCCuDfBk2. Featured Certification.

Beekly, D. L., Ramos, E. M., Lee, W. W., Deitrich, W. D., Jacka, M. E., Wu, J., Hubbard, J. L., Koepsell, T. D., Morris, J. C., Kukull, W. A., et al. The national alzheimer's coordinating center (nacc) database: the uniform data set. *Alzheimer Disease & Associated Disorders*, 21(3): 249–258, 2007.

Brown, J. A., Lee, A. J., Fernhoff, K., Pistone, T., Pasquini, L., Wise, A. B., Staffaroni, A. M., Mandelli, M. L., Lee, S. E., Boxer, A. L., et al. Functional network collapse in neurodegenerative disease. *Nature communications*, 16 (1):10273, 2025.

Bullmore, E. and Sporns, O. Complex brain networks: graph theoretical analysis of structural and functional systems. *Nature Reviews Neuroscience*, 10(3):186–198, Mar 2009. ISSN 1471-003X. doi: 10.1038/nrn2575. URL https://doi.org/10.1038/nrn2575. Published online 4 February 2009.

Calabrese, E., Villanueva-Meyer, J., Rudie, J., Rauschecker, A., Baid, U., Bakas, S., Cha, S., Mongan, J., and Hess, C. The university of california san francisco preoperative diffuse glioma MRI (UCSF-PDGM) (version 5), 2022. URL https://doi.org/10.7937/tcia.bdgf-8v37.

Carmon, J., Heege, J., Necus, J. H., Owen, T. W., Pipa, G., Kaiser, M., Taylor, P. N., and Wang, Y. Reliability and comparability of human brain structural covariance networks. *NeuroImage*, 220:117104, 2020. ISSN 1053-8119. doi: https://doi.org/10.1016/j.neuroimage.2020.117104. URL https://www.sciencedirect.com/science/article/pii/S1053811920305905.

Caro, J. O., de Oliveira Fonseca, A. H., Rizvi, S. A., Rosati, M., Averill, C., Cross, J. L., Mittal, P., Zappala, E., Dhodapkar, R. M., Abdallah, C., et al. Brainlm: A foundation model for brain activity recordings. In *The Twelfth International Conference on Learning Representations*.

Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., and Joulin, A. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9650–9660, 2021.

Chao, D., Anbupalam, T., Jiyang, J., Mather, K. A., Sachdev, P. S., and Wen, W. Brain structural covariances in the ageing brain in the uk biobank. *Brain structure & function*, 229(5):1165–1177, 2024. doi: 10.1007/s00429-024-02794-4.

Chen, S., Ma, K., and Zheng, Y. Med3d: Transfer learning for 3d medical image analysis, 2019. URL https://arxiv.org/abs/1904.00625.

Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. A simple framework for contrastive learning of visual representations. In III, H. D. and Singh, A. (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 1597–1607. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/chen20j.html.

Collins, D. Unbiased nonlinear average age-appropriate brain templates from birth to adulthood. *NeuroImage*, 47: S102, 2009.

Cox, J., Liu, P., Stolte, S. E., Yang, Y., Liu, K., See, K. B., Ju, H., and Fang, R. Brainsegfounder: Towards 3d foundation models for neuroimage segmentation. *Medical Image Analysis*, 97:103301, 2024. ISSN 1361-8415. doi: https://doi.org/10.1016/j.media.2024.103301. URL https://www.sciencedirect.com/science/article/pii/S1361841524002263.

Desikan, R. S., Ségonne, F., Fischl, B., Quinn, B. T., Dickerson, B. C., Blacker, D., Buckner, R. L., Dale, A. M., Maguire, R. P., Hyman, B. T., et al. An automated labeling system for subdividing the human cerebral cortex on mri scans into gyral based regions of interest. *Neuroimage*, 31(3):968–980, 2006.

Dong, Z., Li, R., Wu, Y., Nguyen, T. T., Chong, J., Ji, F., Tong, N., Chen, C., and Zhou, J. H. Brain-jepa: Brain dynamics foundation model with gradient positioning and spatiotemporal masking. *Advances in Neural Information Processing Systems*, 37:86048–86073, 2024.

Doodipala, R. R., Pandey, P., Rojas, C. T., Saikia, M. J., and Sitaram, R. Region-aware reconstruction strategy for pre-training fmri foundation model. In *NeurIPS 2025 Workshop on Foundation Models for the Brain and Body*.

Ghesu, F. C., Georgescu, B., Mansoor, A., Yoo, Y., Neumann, D., Patel, P., Vishwanath, R. S., Balter, J. M., Cao, Y., Grbic, S., et al. Contrastive self-supervised learning from 100 million medical images with optional supervision. *Journal of Medical Imaging*, 64503(1), 2022.

Gong, G., He, Y., Chen, Z. J., and Evans, A. C. Convergence and divergence of thickness correlations with diffusion connections across the human cerebral cortex. *Neuroimage*, 59(2):1239–1248, 2012.

Grill, J.-B., Strub, F., Altché, F., Tallec, C., Richemond, P. H., Buchatskaya, E., Doersch, C., Pires, B. A., Guo, Z. D., Azar, M. G., Piot, B., Kavukcuoglu, K., Munos, R., and Valko, M. Bootstrap your own latent: A new approach to self-supervised learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Grober, E., Lipton, R. B., Sperling, R. A., Papp, K. V., Johnson, K. A., Rentz, D. M., Veroff, A. E., Aisen, P. S., and Ezzati, A. Associations of stages of objective memory impairment with amyloid pet and structural mri: The a4 study. *Neurology*, 98(13):e1327–e1336, 2022.

Gryshchuk, V., Singh, D., Teipel, S., Dyrba, M., and ADNI, AIBL, F. s. g. Contrastive self-supervised learning for neurodegenerative disorder classification. *Frontiers in Neuroinformatics*, 19:1527582, 2025.

Gupta, A., Osman, I., Shehata, M. S., Braun, W. J., and Feldman, R. E. Medmae: A self-supervised backbone for medical imaging tasks. *Computation*, 13(4):88, 2025.

Hatamizadeh, A., Nath, V., Tang, Y., Yang, D., Roth, H. R., and Xu, D. Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images. In Crimi, A. and Bakas, S. (eds.), *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, pp. 272–284, Cham, 2022. Springer International Publishing. ISBN 978-3-031-08999-2.

He, K., Chen, X., Xie, S., Li, Y., Dollár, P., and Girshick, R. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16000–16009, June 2022.

Isensee, F., Schell, M., Pflueger, I., Brugnara, G., Bonekamp, D., Neuberger, U., Wick, A., Schlemmer, H.-P., Heiland, S., Wick, W., Bendszus, M., Maier-Hein, K. H., and Kickingereder, P. Automated brain extraction of multisequence mri using artificial neural networks. *Human Brain Mapping*, 40(17): 4952–4964, 2019. doi: https://doi.org/10.1002/hbm.24750. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/hbm.24750.

IXI Consortium. Ixi dataset. https://brain-development.org/ixi-dataset/, 2004. Brain Development Dataset, Imperial College London and collaborators.

Kaczmarek, E., Szeto, J., Nichyporuk, B., and Arbel, T. Building a general simclr self-supervised foundation model across neurological diseases to advance 3d brain mri diagnoses. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1310–1319, 2025.

Kim, S. Learning general representation of 12-lead electrocardiogram with a joint-embedding predictive architecture. *arXiv preprint arXiv:2410.08559*, 2024.

LaMontagne, P. J., Benzinger, T. L., Morris, J. C., Keefe, S., Hornbeck, R., Xiong, C., Grant, E., Hassenstab, J., Moulder, K., Vlassenko, A. G., et al. Oasis-3: longitudinal neuroimaging, clinical, and cognitive dataset for normal aging and alzheimer disease. *medrxiv*, pp. 2019–12, 2019.

Langley, P. Crafting papers on machine learning. In Langley, P. (ed.), *Proceedings of the 17th International Conference on Machine Learning (ICML 2000)*, pp. 1207–1216, Stanford, CA, 2000. Morgan Kaufmann.

Li, K., Luo, X., Zeng, Q., Liu, X., Li, J., Zhong, S., Zhang, X., Xu, X., Wang, S., Hong, H., Jiaerken, Y., Liu, Z., Zhao, S., Huang, P., Zhang, M., and Chen, Y. Gray matter structural covariance networks patterns associated with autopsy-confirmed late-nc compared to alzheimer's disease pathology.

Neurobiology of Disease, 189:106354, 2023. ISSN 0969-9961. doi: https://doi.org/10.1016/j.nbd.2023.106354. URL https://www.sciencedirect.com/science/article/pii/S0969996123003704.

Li, T. Z., Krishnan, A. R., Zuo, L., Still, J. M., Sandler, K. L., Maldonado, F., Lasko, T. A., and Landman, B. A. Self-supervised learning of imaging and clinical signatures using a multimodal joint-embedding predictive architecture. *arXiv preprint arXiv:2509.15470*, 2025.

Liu, Z., Xu, H., Woicik, A., Shapiro, L. G., Blazes, M., Wu, Y., Steffen, V., Cukras, C. A., Lee, C. S., Zhang, M., Lee, A. Y., and Wang, S. A three-dimensional multi-modal foundation model for optical coherence tomography. *Nature Biomedical Engineering*, 2026. doi: 10.1038/s41551-026-01662-2.

Marek, K., Jennings, D., Lasch, S., Siderowf, A., Tanner, C., Simuni, T., Coffey, C., Kieburtz, K., Flagg, E., Chowdhury, S., Poewe, W., Mollenhauer, B., Klinik, P.-E., Sherer, T., Frasier, M., Meunier, C., Rudolph, A., Casaceli, C., Seibyl, J., Mendick, S., Schuff, N., Zhang, Y., Toga, A., Crawford, K., Ansbach, A., De Blasio, P., Piovella, M., Trojanowski, J., Shaw, L., Singleton, A., Hawkins, K., Eberling, J., Brooks, D., Russell, D., Leary, L., Factor, S., Sommerfeld, B., Hogarth, P., Pighetti, E., Williams, K., Standaert, D., Guthrie, S., Hauser, R., Delgado, H., Jankovic, J., Hunter, C., Stern, M., Tran, B., Leverenz, J., Baca, M., Frank, S., Thomas, C.-A., Richard, I., Deeley, C., Rees, L., Sprenger, F., Lang, E., Shill, H., Obradov, S., Fernandez, H., Winters, A., Berg, D., Gauss, K., Galasko, D., Fontaine, D., Mari, Z., Gerstenhaber, M., Brooks, D., Malloy, S., Barone, P., Longo, K., Comery, T., Ravina, B., Grachev, I., Gallagher, K., Collins, M., Widnell, K. L., Ostrowizki, S., Fontoura, P., Ho, T., Luthman, J., van der Brug, M., Reith, A. D., and Taylor, P. The parkinson progression marker initiative (ppmi). *Progress in Neurobiology*, 95(4):629–635, 2011. ISSN 0301-0082. doi: https://doi.org/10.1016/j.pneurobio.2011.09.005. URL https://www.sciencedirect.com/science/article/pii/S0301008211001651. Biological Markers for Neurodegenerative Diseases.

Mazher, M., Parker, G. J. M., and Alexander, D. C. Towards generalisable foundation models for brain mri, 2025. URL https://arxiv.org/abs/2510.23415.

Mongay-Ochoa, N., Gonzalez-Escamilla, G., Fleischer, V., Pareto, D., Rovira, À., Sastre-Garriga, J., and Groppa, S. Structural covariance analysis for neurodegenerative and neuroinflammatory brain disorders. *Brain*, pp. awaf151, 2025.

Munk, A., Ambsdorf, J., Llambias, S., and Nielsen, M. Amaes: Augmented masked autoencoder pretraining on public brain mri data for 3d-native segmentation. *arXiv preprint arXiv:2408.00640*, 2024.

Nigro, S., Filardi, M., Tafuri, B., De Blasi, R., Dell'Abate, M. T., Giugno, A., Gnoni, V., Milella, G., Urso, D., Zecca, C., Zoccolella, S., and Logroscino, G. Radiomics feature similarity: A novel approach for characterizing brain network changes in patients with behavioral variant frontotemporal dementia. *NeuroImage: Clinical*, 46:103780, 2025. doi: 10.1016/j.nicl.2025.103780.

Petersen, R. C., Aisen, P. S., Beckett, L. A., Donohue, M. C., Gamst, A. C., Harvey, D. J., Jack Jr, C., Jagust, W. J., Shaw, L. M., Toga, A. W., et al. Alzheimer's disease neuroimaging initiative (adni) clinical characterization. *Neurology*, 74(3):201–209, 2010.

Qiu, J., Wu, J., Wei, H., Shi, P., Zhang, M., Sun, Y., Li, L., Liu, H., Liu, H., Hou, S., Zhao, Y., Shi, X., Xian, J., Qu, X., Zhu, S., Pan, L., Chen, X., Zhang, X., Jiang, S., Wang, K., Yang, C., Chen, M., Fan, S., Hu, J., Lv, A., Miao, H., Guo, L., Zhang, S., Pei, C., Fan, X., Lei, J., Wei, T., Duan, J., Liu, C., Xia, X., Xiong, S., Li, J., Lam, K., Lo, B., Tham, Y. C., Wong, T. Y., Wang, N., and Yuan, W. Development and validation of a multimodal multitask vision foundation model for generalist ophthalmic artificial intelligence. *NEJM AI*, 1(12):AIoa2300221, 2024. doi: 10.1056/AIoa2300221. URL https://ai.nejm.org/doi/full/10.1056/AIoa2300221.

Roberts, R. O., Geda, Y. E., Knopman, D. S., Cha, R. H., Pankratz, V. S., Boeve, B. F., Ivnik, R. J., Tangalos, E. G., Petersen, R. C., and Rocca, W. A. The mayo clinic study of aging: Design and sampling, participation, baseline measures and sample characteristics. *Neuroepidemiology*, 30(1):58–69, 02 2008. ISSN 0251-5350. doi: 10.1159/000115751. URL https://doi.org/10.1159/000115751.

Rohrer, J. D. and Rosen, H. J. Neuroimaging in frontotemporal dementia. *International Review of Psychiatry*, 25 (2):221–229, 2013.

Rorden, C. and Absher, J. Stroke outcome optimization project (soop), 2023. URL https://doi.org/10.18112/openneuro.ds004889.v1.0.0. BIDS Version 1.6.0, CC0 license.

Rushmore, R. J., Sunderland, K., Carrington, H., Chen, J., Halle, M., Lasso, A., Rizzoni, E., Vessey, B., et al. Anatomically curated segmentation of human subcortical structures in high resolution magnetic resonance imaging: An open science approach. *Frontiers in Neuroanatomy*, 16:894606, 2022.

Seeley, B. Neurodegenerative diseases target large-scale human brain networks. *Alzheimer's & Dementia*, 6(4): S167, 2010.

Shakouri, M., Iranmanesh, F., and Eftekhari, M. Dino-cxr: A self supervised method based on vision transformer for chest x-ray classification. In Bebis, G., Ghiasi, G., Fang, Y., Sharf, A., Dong, Y., Weaver, C., Leo, Z., Jr., J. J. L., and Kohli, L. (eds.), *Advances in Visual Computing*, pp. 320–331. Springer Nature Switzerland, 2023. ISBN 978-3-031-47966-3. doi: 10.1007/978-3-031-47966-3_25.

Silva-Rodríguez, J., Chakor, H., Kobbi, R., Dolz, J., and Ben Ayed, I. A foundation language-image model of the retina (flair): encoding expert knowledge in text supervision. *Medical Image Analysis*, 99:103357, 2025. ISSN 1361-8415. doi: https://doi.org/10.1016/j.media.2024.103357. URL https://www.sciencedirect.com/science/article/pii/S1361841524002822.

Tak, D., Garomsa, B. A., Zapaishchykova, A., Chaunzwa, T. L., Climent Pardo, J. C., Ye, Z., Zielke, J., Ravipati, Y., Pai, S., Vajapeyam, S., Mahootiha, M., Parker, M., Pike, L. R. G., Smith, C., Familiar, A. M., Liu, K. X., Prabhu, S., Arnaout, O., Bandopadhayay, P., Nabavizadeh, A., Mueller, S., Aerts, H. J., Huang, R. Y., Poussaint, T. Y., and Kann, B. H. A generalizable foundation model for analysis of human brain MRI. *Nature Neuroscience*, 29 (4):945–956, 2026. ISSN 1546-1726. doi: 10.1038/s41593-026-02202-6. URL https://doi.org/10.1038/s41593-026-02202-6.

Tang, Y., Yang, D., Li, W., Roth, H. R., Landman, B., Xu, D., Nath, V., and Hatamizadeh, A. Self-supervised pre-training of swin transformers for 3d medical image analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20730–20740, June 2022.

Thomas, A., Ré, C., and Poldrack, R. Self-supervised learning of brain dynamics from broad neuroimaging data. *Advances in neural information processing systems*, 35: 21255–21269, 2022.

Tong, Z., Song, Y., Wang, J., and Wang, L. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. In Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., and Oh, A. (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 10078–10093. Curran Associates, Inc., 2022.

Valanarasu, J. M. J., Tang, Y., Yang, D., Xu, Z., Zhao, C., Li, W., Patel, V. M., Landman, B. A., Xu, D., He, Y., and Nath, V. Disruptive autoencoders: Leveraging low-level features for 3d medical image pre-training. In Burgos, N., Petitjean, C., Vakalopoulou, M., Christodoulidis, S., Coupe, P., Delingette, H., Lartizien, C., and Mateus, D. (eds.), *Proceedings of The 7nd International Conference on Medical Imaging with Deep Learning*, volume 250 of *Proceedings of Machine Learning Research*, pp. 1553–1570. PMLR, 03–05 Jul 2024. URL https://proceedings.mlr.press/v250/valanarasu24a.html.

Van Essen, D. C., Smith, S. M., Barch, D. M., Behrens, T. E., Yacoub, E., and Ugurbil, K. The wu-minn human connectome project: An overview. *NeuroImage*, 80:62–79, 2013. ISSN 1053-8119. doi: https://doi.org/10.1016/j.neuroimage.2013.05.041. URL https://www.sciencedirect.com/science/article/pii/S1053811913005351. Mapping the Connectome.

Vogel, J. W., Corriveau-Lecavalier, N., Franzmeier, N., Pereira, J. B., Brown, J. A., Maass, A., Botha, H., Seeley, W. W., Bassett, D. S., Jones, D. T., et al. Connectome-based modelling of neurodegenerative diseases: towards precision medicine and mechanistic insight. *Nature Reviews Neuroscience*, 24(10):620–639, 2023.

Wang, S., Safari, M., Li, Q., Chang, C.-W., LJ Qiu, R., Roper, J., Yu, D. S., and Yang, X. Vision foundation model for 3d magnetic resonance imaging segmentation, classification, and registration. *Medical Image Analysis*, 110:103992, 2026. ISSN 1361-8415. doi: https://doi.org/10.1016/j.media.2026.103992. URL https://www.sciencedirect.com/science/article/pii/S1361841526000617.

Wei, D., Zhuang, K., Ai, L., Chen, Q., Yang, W., Liu, W., Wang, K., Sun, J., and Qiu, J. Structural and functional brain scans from the cross-sectional southwest university adult lifespan dataset. *Scientific data*, 5(1):180134, 2018.

Wheeler, T., Hunter, K., Garcia, P., Li, H., Thomson, A., Hunter, A., and Mehanian, C. Self-supervised contrastive learning improves machine learning discrimination of full thickness macular holes from epiretinal membranes in retinal oct scans. *medRxiv*, pp. 2023–11, 2023.

Xie, Y., Gu, L., Harada, T., Zhang, J., Xia, Y., and Wu, Q. Rethinking masked image modelling for medical image representation. *Medical Image Analysis*, 98:103304, 2024.

Xu, T., Hosseini, S., Anderson, C., Rinaldi, A., Krishnan, R. G., Martel, A. L., and Goubran, M. A generalizable 3d framework and model for self-supervised learning in medical imaging. *npj Digital Medicine*, 8(1):639, 2025.

Xu, Y., Raj, A., Victor, J., and Initiative, A. D. N. Systematic differences between visually-relevant global and local image statistics of brain mri and natural scenes. *bioRxiv*, pp. 271338, 2018.

Yang, G., Du, K., Yang, Z., Du, Y., Cheung, E. Y. W., Zheng, Y., Yang, M., Kourtzi, Z., Schönlieb, C.-B., Wang, S., and

Initiative, A. D. N. Adfound: A foundation model for diagnosis and prognosis of alzheimer's disease. *IEEE Journal of Biomedical and Health Informatics*, 29(11): 8395–8408, 2025. doi: 10.1109/JBHI.2025.3576436.

Yu, K., Zhou, Y., Bai, Y., Soh, Z. D., Xu, X., Goh, R. S. M., Cheng, C.-Y., and Liu, Y. UrFound: Towards Universal Retinal Foundation Models via Knowledge-Guided Masked Modeling . In *proceedings of Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, volume LNCS 15012. Springer Nature Switzerland, October 2024.

Zbontar, J., Jing, L., Misra, I., LeCun, Y., and Deny, S. Barlow twins: Self-supervised learning via redundancy reduction. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, pp. 12310–12320. PMLR, 2021.

Zhao, K., Zheng, Q., Che, T., Martin, D., Li, Q., Ding, Y., Zheng, Y., Liu, Y., and Li, S. Regional radiomics similarity networks (r2sn) in the human brain: reproducibility, small-world and biological basis. *bioRxiv*, 2020. doi: 10.1101/2020.12.09.418509.

Zhou, L., Liu, H., Bae, J., He, J., Samaras, D., and Prasanna, P. Self pre-training with masked autoencoders for medical image classification and segmentation. In *2023 IEEE 20th international symposium on biomedical imaging (ISBI)*, pp. 1–6. IEEE, 2023a.

Zhou, Y., Chia, M. A., Wagner, S. K., Ayhan, M. S., Williamson, D. J., Struyven, R. R., Liu, T., Xu, M., Lozano, M. G., Woodward-Court, P., Kihara, Y., Allen, N., Gallacher, J. E., Littlejohns, T., Aslam, T., Bishop, P., Black, G., Sergouniotis, P., Atan, D., Dick, A. D., Williams, C., Barman, S., Barrett, J. H., Mackie, S., Braithwaite, T., Carare, R. O., Ennis, S., Gibson, J., Lotery, A. J., Self, J., Chakravarthy, U., Hogg, R. E., Paterson, E., Woodside, J., Peto, T., Mckay, G., Mcguinness, B., Foster, P. J., Balaskas, K., Khawaja, A. P., Pontikos, N., Rahi, J. S., Lascaratos, G., Patel, P. J., Chan, M., Chua, S. Y., Day, A., Desai, P., Egan, C., Fruttiger, M., Garway-Heath, D. F., Hardcastle, A., Khaw, S. P. T., Moore, T., Sivaprasad, S., Strouthidis, N., Thomas, D., Tufail, A., Viswanathan, A. C., Dhillon, B., Macgillivray, T., Sudlow, C., Vitart, V., Doney, A., Trucco, E., Guggeinheim, J. A., Morgan, J. E., Hammond, C. J., Williams, K., Hysi, P., Harding, S. P., Zheng, Y., Luben, R., Luthert, P., Sun, Z., McKibbin, M., O'Sullivan, E., Oram, R., Weedon, M., Owen, C. G., Rudnicka, A. R., Sattar, N., Steel, D., Stratton, I., Tapp, R., Yates, M. M., Petzold, A., Madhusudhan, S., Altmann, A., Lee, A. Y., Topol, E. J., Denniston, A. K., Alexander, D. C., and Keane, P. A. A foundation model for generalizable disease detection from retinal images. *Nature 2023 622:7981*, 622: 156–163, 9 2023b. ISSN 1476-4687. doi: 10.1038/s41586-023-06555-x. URL https://www.nature.com/articles/s41586-023-06555-x.

Zhuang, J., Luo, L., Wang, Q., Wu, M., Luo, L., and Chen, H. Advancing volumetric medical image segmentation via global-local masked autoencoders. *IEEE Transactions on Medical Imaging*, 2025a.

Zhuang, J., Wu, L., Wang, Q., Fei, P., Vardhanabhuti, V., Luo, L., and Chen, H. Mim: Mask in mask self-supervised pre-training for 3d medical image analysis. *IEEE Transactions on Medical Imaging*, 44(9):3727–3740, 2025b. doi: 10.1109/TMI.2025.3564382.

A. Related Work

Self-supervised learning (SSL) has emerged as a powerful paradigm for leveraging large-scale unlabeled data in biomedical imaging, where expert annotations are costly and scarce. In the broader computer vision literature, SSL methods are typically grouped by their core supervisory principle: contrastive learning (which enforces invariance across augmented or paired views) (Chen et al., 2020), reconstruction-based approaches (which recover masked or corrupted input signals at the pixel or voxel level) (He et al., 2022; Tong et al., 2022), latent predictive architectures (which forecast representations of unseen parts in a non-generative student-teacher setup) (Assran et al., 2023; Bardes et al., 2024), and self-distillation (which aligns predictions between a momentum-updated teacher and student network) (Caron et al., 2021; Grill et al., 2020). These general-domain formulations, originally developed for natural images, have been widely adapted to medical data.

As summarized in Table 3, biomedical SSL efforts largely mirror this taxonomy. Contrastive methods follow the standard instance-discrimination objective of pulling representations of positive pairs, typically augmented views of the same sample, closer while repelling negatives. Examples include BrainIAC, which applies SimCLR-style contrastive pretraining across multiparametric brain MRI (Tak et al., 2026), NNCLR-based approaches on structural MRI slices to derive disease-relevant anatomical embeddings (Gryshchuk et al., 2025), and RaSCL, which constructs positives from neighboring OCT slices within an otherwise conventional contrastive framework (Wheeler et al., 2023). Large-scale efforts such as contrastive pretraining on over 100 million multimodal radiology images seek modality-invariant representations (Ghesu et al., 2022). Vision-language contrastive approaches have also been explored for retinal imaging; for example, FLAIR (Silva-Rodríguez et al., 2025) trains a CLIP-style model on fundus images by aligning image embeddings with descriptive textual prompts that encode expert domain knowledge (fine-grained pathological features and hierarchical relationships between conditions). Even anatomically motivated variants like AnatCL remain anchored in the standard contrastive loss structure; they incorporate age or ROI-derived measures only as inter-image weighting factors rather than redefining the supervisory signal itself (Barbano et al., 2026).

Reconstruction-based SSL, commonly instantiated via masked autoencoding, relies on explicit recovery of masked input regions from visible context. Multiple works apply this principle directly to medical images: MedMAE (Gupta et al., 2025), GL-MAE (Zhuang et al., 2025a), MedIM (Xie et al., 2024), and SelfMedMAE (Zhou et al., 2023a) perform standard patch- or region-level reconstruction across modalities. RetFound demonstrates its utility for retinal imaging by pretraining a masked autoencoder before label-efficient fine-tuning (Zhou et al., 2023b). Recent efforts further advance this paradigm by incorporating domain knowledge. UrFound (Yu et al., 2024) proposes a knowledge-guided masked modeling strategy on multimodal CFP and OCT images ($\sim$180k), combining masked image reconstruction with conditional masked language modeling to integrate expert annotations as text supervision. Domain adaptations often refine the masking strategy, for example ROI-guided masking with the AAL3 atlas in fMRI volumes (Doodipala et al.), or spatiotemporal masking in Brain Dynamics SSL (Thomas et al., 2022) and BrainLM (Caro et al.). Hierarchical extensions such as MiM (Zhuang et al., 2025b) and Triad (Wang et al., 2026), as well as augmentation-aware variants like AMAES (Munk et al., 2024), similarly operate within the reconstruction paradigm while adding cross-scale alignment or equivariance constraints as auxiliary components.

Latent predictive approaches, exemplified by joint-embedding predictive architectures (JEPA), shift supervision to the representation space: a student network predicts the latent embeddings of masked regions produced by a teacher, avoiding both raw-input reconstruction and explicit contrastive negatives. Brain-JEPA applies this to spatiotemporal fMRI blocks with brain-specific positioning as an inductive bias (Dong et al., 2024). Multimodal extensions handle paired CT and EHR data while retaining the JEPA-style prediction objective (Li et al., 2025), and ECG-JEPA predicts missing latent patches in 12-lead waveforms (Kim, 2024).

Self-distillation methods transfer teacher-student alignment objectives, as in DINO, to medical imaging. DINO-CXR adapts it for chest X-rays with suitable augmentations and backbones (Shakouri et al., 2023). Volumetric extensions such as 3DINO-ViT (Xu et al., 2025) and BrainFound (Mazher et al., 2025) process 3D MRI or CT volumes, often via slice-wise strategies, while keeping the core alignment of sharpened student predictions to teacher outputs unchanged. Domain adaptation occurs primarily through data handling and architecture.

Hybrid approaches combine multiple standard objectives, such as masked reconstruction with contrastive alignment (ALICE (Kaczmarek et al., 2025), ADFound (Yang et al., 2025)) or additional pretext tasks (SwinUNETR (Tang et al., 2022)). Recent work has extended this to 3D volumetric data with multi-modal integration; for instance, OCTCube-M employs a multi-modal contrastive learning method (COEP) to pretrain a 3D foundation model on over 26,000 OCT volumes paired with other retinal imaging modalities such as infrared and en face (EF) images, achieving strong performance on retinal disease prediction and geographic atrophy prognosis while emphasizing cross-modality alignment (Liu et al., 2026).

Collectively, these works reveal a consistent pattern. Despite diversity in modalities, scales, and downstream applications, biomedical SSL has predominantly imported general-purpose objectives from natural image pretraining: contrastive invariance, masked reconstruction, latent prediction, or self-distillation. Anatomical or clinical knowledge enters mainly through auxiliary mechanisms such as modality-specific augmentations, structured masking policies, metadata weighting, architectural modifications (for example, 3D or slice-wise ViTs), or explicit hybrid losses that still optimize combinations of standard components. This stands in contrast to clinical neuroimaging practice, in which experts routinely interpret inter-regional dependencies, hemispheric symmetry, and structural covariance as diagnostically meaningful signals.

What has been largely missing is a supervisory principle that directly embeds the intrinsic biological organization of the data, such as brain structural covariance, as the primary learning signal rather than an add-on. Our work addresses this gap by introducing a neuroanatomy-informed SSL task centered on preservation of inter-regional and inter-hemispheric structural relationships across samples. By making domain-structured reasoning the core of pretraining, we move beyond adaptation of general-purpose SSL toward objectives better aligned with the underlying biology of brain MRI.

The proposed formulation is fundamentally different from Barlow Twins (Zbontar et al., 2021), where covariance is modeled on latent embedding dimensions to learn general-purpose, view-invariant and redundancy-reduced representation. In our proposed formulation, covariance is computed between anatomically grounded brain regions, and supervision is restricted to biologically meaningful homologous region pairs rather than latent embedding dimensions. Consequently, our objective preserves structured neuroanatomical relationships across samples for structural brain MRI, instead of enforcing generic decorrelation or redundancy reduction in the latent representation space.

Table 3. Summary of self-supervised learning (SSL) approaches in biomedicine grouped by primary supervisory principle.

Method	**Description**
Contrastive SSL: The core objective is invariance across (augmented or paired) views; representations of positive pairs are pulled closer while negatives are pushed apart using standard contrastive losses (InfoNCE / NT-Xent or direct variants).	
BrainIAC (Tak et al., 2026)	Uses SimCLR-style contrastive learning to learn invariant representations across multi-parametric brain MRI augmentations.
NNCLR (Gryshchuk et al., 2025)	Applies nearest-neighbour contrastive learning to structural MRI slices to learn disease-relevant anatomical embeddings.
RaSCL (OCT) (Wheeler et al., 2023)	Uses standard contrastive learning but constructs positive pairs from neighbouring OCT slices to enforce retinal anatomical consistency.
Ghesu et al. (2022) (100M Radiology)	Performs large-scale contrastive learning across 100M multimodal radiology images to obtain modality-invariant representations.
AnatCL (Barbano et al., 2026)	Extends weakly contrastive learning by weighting similarity using age and ROI-derived anatomical measures; keeps the contrastive-learning loss structure (degree-of-positiveness / kernel-weighted attraction), i.e., an inter-image weighting prior inside a standard contrastive framework.
FLIAR (Silva-Rodríguez et al., 2025)	FLAIR extends standard vision–language contrastive learning by introducing expert-designed textual descriptions of retinal pathologies (e.g., lesion features and disease hierarchies) instead of simple class labels, enabling richer image–text alignment while preserving the conventional CLIP-style contrastive objective and optimization scheme.
Reconstruction-based SSL: The supervisory signal is explicit pixel-/voxel-/signal-level reconstruction of masked input regions/tokens from visible context (standard MAE / BERT-style objective).	
MedMAE (Gupta et al., 2025)	Standard masked autoencoding applied to medical images; the model reconstructs randomly masked patches/regions through direct pixel/voxel recovery.

Continued on next page…

Table 3 continued...

Method	Description
GL-MAE (Zhuang et al., 2025a)	
MedIM (Xie et al., 2024)	
SelfMedMAE (Zhou et al., 2023a)	
ROI-guided MAE (fMRI) (Doodipala et al.)	Extends standard MAE with an ROI-guided masking strategy (AAL3 atlas) that selectively masks entire anatomically coherent brain regions in 4D fMRI volumes; atlas-defined ROIs choose semantically coherent masks within a standard masked-reconstruction SSL objective.
Brain Dynamics SSL (Thomas et al., 2022)	Uses masked + causal sequence autoencoding on parcellated fMRI time-series to reconstruct missing brain-network activity.
BrainLM (Caro et al.)	Pretrains a Transformer via spatiotemporal masking and reconstruction of masked fMRI segments; training objective is to predict masked content from context (MAE family).
MiM (Zhuang et al., 2025b)	Hierarchical masked autoencoding framework that performs multi-level reconstruction of 3D medical volumes with cross-level alignment to capture anatomical structures across scales, pretrained on large-scale CT datasets and applied to downstream tasks such as organ/tumor segmentation and disease classification.
Triad (Wang et al., 2026)	Masked image modeling-based 3D MRI foundation model using SimMIM reconstruction with auxiliary text-guided modality distribution alignment, pretrained on large-scale MRI datasets and applied to segmentation, classification, and registration tasks.
AMAES (Munk et al., 2024)	Applies masked autoencoding with augmentation-aware reconstruction, learning representations invariant to intensity perturbations and equivariant to spatial transformations, while retaining a standard reconstruction-based SSL objective.
RETFound (Zhou et al., 2023b)	Builds a retinal foundation model using MAE-based self-supervised pretraining on 1.6 million unlabeled CFP and OCT images, enabling label-efficient transfer and strong generalization across diverse ocular diagnosis, prognosis, and systemic disease prediction tasks.
UrFound (Yu et al., 2024)	Introduces knowledge-guided masked modeling by combining masked image reconstruction with conditional masked language modeling of expert clinical descriptions, enabling multimodal retinal foundation learning across multimodal CFP and OCT images ($\sim$180k) while retaining a standard reconstruction-based self-supervised learning paradigm.
VisionFM (Qiu et al., 2024)	A multi-modal, multi-task ophthalmic foundation model pretrained via self-supervised masked reconstruction on 3.4 million images across 8 modalities (fundus, OCT, FFA, slit-lamp, UBM, etc.); learns generalist representations that support disease diagnosis, segmentation, landmark detection, prognosis, and systemic biomarker prediction with strong generalization to new modalities and devices.

Latent-prediction / JEPA-style SSL: The objective is prediction of latent representations (not raw input) of masked regions in a student-teacher setup; explicitly non-generative and non-contrastive (standard JEPA framework).

Continued on next page...

Table 3 continued...

Method	Description
Brain-JEPA (Dong et al., 2024)	Predicts latent representations of masked spatiotemporal fMRI blocks from observed context rather than reconstructing raw signals; brain-specific positioning/masking acts as an inductive bias on top of the standard JEPA formulation.
Multimodal-JEPA (CT+EHR) (Li et al., 2025)	Uses a joint-embedding predictive architecture to pretrain on paired CT + EHR from unlabeled archives, retaining JEPA-style latent prediction while changing the modality mix and data source.
ECG-JEPA (Kim, 2024)	Learns 12-lead ECG representations by predicting missing latent patches rather than reconstructing waveforms; supervision is representation prediction over masked regions.
Self-distillation SSL (Teacher–Student / DINO-style): The supervisory signal is knowledge distillation between a momentum teacher and a student network over augmented views of the same input; no explicit contrastive negatives or reconstruction target, and representations are learned by aligning student predictions to teacher outputs via cross-entropy on softmax-sharpened embeddings.	
DINO-CXR (Shakouri et al., 2023)	Self-distillation-based SSL framework adapting DINO with a vision transformer backbone for chest X-ray imaging, learning representations via teacher–student alignment and applied to pneumonia and COVID-19 detection tasks.
3DINO-ViT (Xu et al., 2025)	Self-distillation-based SSL framework adapting DINOv2 to 3D medical imaging by aligning student–teacher representations at both image and patch levels; pretrained on $\sim$100,000 multimodal MRI and CT volumes across 10+ organs, yielding general-purpose representations transferable to segmentation, classification, and out-of-distribution tasks.
BrainFound (Mazher et al., 2025)	Extends DINOv2 to volumetric brain MRI via a slice-wise strategy, converting 3D volumes into stacked axial slices (T1, T2, FLAIR as channels) and continuing the DINOv2 teacher–student self-distillation loss unchanged on $\sim$10,000 unlabeled brain MRI volumes from 12 public datasets.
Hybrid SSL (Contrastive + Reconstruction): Training jointly optimises reconstruction and contrastive alignment.	
ALICE (Kaczmarek et al., 2025)	Combines masked image modeling with contrastive learning and anatomical semantic alignment/pair crafting; optimization is an explicit sum of standard contrastive and masked-reconstruction objectives.
ADFound (Yang et al., 2025)	Pretrains on multimodal neuroimaging using a multimodal MAE reconstruction loss plus dual contrastive learning; learning signal is explicitly the joint reconstruction + contrastive objectives.
SwinUNETR (Tang et al., 2022; Hatamizadeh et al., 2022)	Transformer-based SSL framework combining masked volume inpainting, contrastive learning, and rotation prediction for 3D medical images, pretrained on large-scale unlabeled CT data and applied to downstream tasks such as multi-organ segmentation (e.g., BTCV, MSD) and brain tumor segmentation on MRI (BraTS).
OCTCube / OCTCube-M (Liu et al., 2026)	Combines 3D masked autoencoding for volumetric OCT reconstruction with contrastive cross-modal embedding pretraining (COEP) to align OCT with infrared and fundus autofluorescence retinal images, preserving standard reconstruction and contrastive SSL objectives while enabling multimodal foundation learning.

Continued on next page...

Table 3 continued...

Method	Description
Neuroanatomy-informed / domain-structured SSL: Domain-specific neuroanatomical priors explicitly guide representation learning.	
NAI-SSL **(ours)**	Introduces a brain structural covariance pretext task that explicitly enforces preservation of inter-regional and inter-hemispheric structural relationships during representation learning.

B. Self-Supervised Task Design Choices and Implementation

The proposed self-supervised task involves several design choices that affect the quality of learned representations. We discuss key design components, which is motivated by a combination of statistical considerations and neuroanatomical priors, and we analyze their impact empirically in Sec. F.

Region-level feature design. A key design choice is the feature of each ROI. In neuroimaging, inter-hemispheric structural covariance is traditionally studied using region-level *volumetric features*, such as cortical thickness and gray matter volume (Alexander-Bloch et al., 2013; Xu et al., 2018). These features are derived from anatomical segmentation and provide biologically meaningful descriptors of regional morphology. However, directly adopting volumetric features in our self-supervised task is impractical, as they require accurate segmentation for each sample. While this can be performed as offline preprocessing for input data $\mathbf{x}$, applying segmentation to model output $\mathbf{y}$ during training would incur significant computational overhead, making large-scale self-supervised pretraining infeasible.

To address this limitation, we instead explore *radiomics-based features* derived from voxel intensity within each ROI. Recent neuroimaging studies have increasingly adopted such features (Zhao et al., 2020; Nigro et al., 2025), including mean intensity and higher-order statistical descriptors, to characterize structural variations. In this work, we consider scalar descriptors such as mean intensity, standard deviation, and entropy. *Importantly, this design avoids repeated anatomical segmentation during training.* Radiomics features for model output $\mathbf{y}$ can be computed directly using the ROI definitions obtained from the input data $\mathbf{x}$, enabling efficient and scalable self-supervised learning. We empirically evaluate different feature choices in Sec. F and show that simple intensity-based descriptors provide a favorable trade-off between computational efficiency and representation quality, while remaining sufficient to capture meaningful inter-regional relationships.

Minibatch-based covariance estimation. The proposed task relies on covariance computed across samples, which in classical neuroimaging is defined over the full data distribution (Alexander-Bloch et al., 2013). In our self-supervised formulation, we instead estimate this quantity using minibatches during training. Given a minibatch of size B, the covariance matrix computed from the batch provides a stochastic estimate of structural relationships across samples.

This design is motivated by the characteristics of large-scale biomedical self-supervised pretraining (Zhou et al., 2023b; Tak et al., 2026). While a population-level covariance estimate can be obtained using all training data (e.g., via moving averages), such an approach may be suboptimal in practice. Pretraining datasets are often composed of multiple cohorts with heterogeneous acquisition protocols, demographics, and health conditions (Zhou et al., 2023b; Tak et al., 2026). A single global covariance estimate would aggregate these diverse sources and may obscure meaningful variations in structural relationships.

In contrast, minibatch-based covariance estimation provides a data-adaptive alternative that captures structural relationships within subsets of samples (Supp H). When the minibatch size is sufficiently large to ensure statistical stability, this approach enables the model to learn representations that are robust across heterogeneous data distributions while remaining sensitive to cohort-specific variation. We further analyze the effect of batch size and empirically verify that the chosen batch size yields stable covariance estimates (see Sec. F).

Why homologous region pairs. The proposed task enforces covariance consistency between homologous regions across the two hemispheres, rather than using the full inter-regional covariance matrix. From a neuroanatomical perspective, homologous regions are anatomically matched and exhibit strong inter-hemispheric structural consistency, representing some of the most reliable forms of structural dependency in the brain (Alexander-Bloch et al., 2013; Xu et al., 2018). Focusing on

these pairs, therefore, provides a biologically grounded signal that reflects established organizational principles. This design also reduces the dimensionality of the supervision, further improving stability in minibatch-based estimation.

Implementation. The proposed formulation defines supervision at the level of statistics derived across a group of samples, rather than on individual outputs, and is not specific to a particular architectural choice. In our implementation, the input is first perturbed (e.g., by masking or noise corruption) and then processed by the model. We instantiate the model as an encoder followed by a prediction head. The prediction head produces outputs in the input (image) domain, from which region-level features are computed. These features, collected across samples within a minibatch, are used to compute inter-regional covariance $\mathbf{C_y}$, which is matched to the covariance $\mathbf{C_x}$ computed from the input. *Importantly, the outputs of the prediction head are not directly supervised; supervision is defined through cross-sample statistics derived from them.* Following standard SSL practice, the prediction head is used only during pretraining and is discarded afterward, with the encoder retained for downstream tasks.

C. Dataset Description

Table 4. Pre-training dataset count

DATA SET	COUNT
A4 (GROBER ET AL., 2022)	1819
MCSA (ROBERTS ET AL., 2008)	3095
NACC (BEEKLY ET AL., 2007)	1626
OASIS (LAMONTAGNE ET AL., 2019)	424
PPMI (MAREK ET AL., 2011)	2401
SOOP (RORDEN & ABSHER, 2023)	1393
WU1200 (VAN ESSEN ET AL., 2013)	1783
TOTAL	12,541

Table 5. Fine-tuning and evaluation dataset statistics

	DATASET	TASK	TOTAL
	ADNI (PETERSEN ET AL., 2010)	AD/MCI/CN CLASSIFICATION	2549
		MCI/CN CLASSIFICATION	1969
	IXI (IXI CONSORTIUM, 2004)	AGE PREDICTION AND GENDER CLASSIFICATION	562
	NIFD (ROHRER & ROSEN, 2013)	FTD CLASSIFICATION	1466
	UCSF-PDGM (CALABRESE ET AL., 2022)	TUMOUR GRADE	466
EXTERNAL	OASIS-3 (LAMONTAGNE ET AL., 2019)	AD/MCI/CN CLASSIFICATION	251
		MCI/CN CLASSIFICATION	230
	SALD (WEI ET AL., 2018)	AGE PREDICTION	434
		GENDER CLASSIFICATION	494

C.1. Data Pre-processing

All datasets were processed through standardised preprocessing steps to ensure data consistency and quality control across all structural magnetic resonance imaging (MRI) scans. Raw DICOM images were first converted to NIfTI format using the `dcm2nii` package. To correct for low-frequency intensity inhomogeneities inherent to MRI acquisition, N4 bias field correction was applied using `SimpleITK`. All images were then skull stripped using the `HD-BET` framework (Isensee et al., 2019) and finally rigidly registered to the MNI (Collins, 2009) standard brain space.

D. Experiment Details

Downstream task adaptation. For downstream evaluation, we closely follow the BrainIAC (Tak et al., 2026) evaluation protocol. Representation quality is evaluated using the standard SSL linear probing protocol across **eight** downstream evaluation settings, spanning clinical classification, demographic prediction, and external benchmark tasks, including: MCI vs. CN classification, AD vs. MCI vs. CN classification, FTD vs. CN classification, brain age prediction (regression), gender classification, and tumor grading. *Importantly, the pretraining dataset does not contain samples from any downstream evaluation tasks, ensuring strict separation between pretraining and downstream evaluation.*

Evaluation metrics. Classification performance is measured using balanced accuracy, AUC, and F1 score, while regression performance is evaluated using mean absolute error (shown as L1) between predicted and chronological age. Results are reported on 20% of held-out test sets.

Pretraining configuration. We compare the proposed NAI-SSL against representative state-of-the-art (SOTA) SSL approaches spanning major SSL paradigms, including contrastive learning and reconstruction-based SSL. For contrastive SSL, we adopt SimCLR, which has been used in SOTA BrainIAC (Tak et al., 2026) for structural brain MRI. For reconstruction-based SSL, we follow SOTA MiM (Zhuang et al., 2025b) and adopt MAE. Masking is chosen over noise-corruption as perturbation in NAI-SSL based on ablation results. All the SSL methods are pretrained for 50 epochs.

E. Pretraining Details

Table 6. NAISSL pretraining hyperparameters.

Category	NAISSL
Data	
Dataset	Multi-site T1-weighted MRI
Input resolution	96x96x96
Patch size	16x16x16
Sampling rate	1
Batch size	64
Model	
Backbone	ViT-B/16
Decoder depth	8
Decoder embed dim	576
Mask tokens	Enabled
Masking	
Mask type	random patch masking
Spatial scale	0.6
Temporal scale	1.0
Optimization	
Epochs	50
Optimizer	AdamW
Base learning rate	1.0×10^{-6}
Warm-up epochs	10
Weight decay	0.05

In this section, we report our NAI-SSL pretraining details. Tab. 6 summarizes the main hyperparameters used during pretraining.

Table 7. Ablation study on (a) batch size, (b) pretext radiomic features, (c) brain atlas, (d) covariance estimation strategy The default configuration (batch size = 64, Mean+Var features, Harvard-Oxford atlas, minibatch-level covariance estimation) is marked with †. The best result per column within each section is **bolded**. BAcc = balanced accuracy (%). ↑: higher is better; ↓: lower is better.

	AD/MCI/CN (ext)			Age	Gender			MCI/CN			FTD			Tumor Grade		
Config	BAcc↑	AUC↑	F1↑	L1↓	BAcc↑	AUC↑	F1↑	BAcc↑	AUC↑	F1↑	BAcc↑	AUC↑	F1↑	BAcc↑	AUC↑	F1↑
(a) Batch size																
bs = 32	51.19	0.7805	0.5147	9.2622	82.84	0.9048	0.8313	74.20	0.8137	0.7540	82.57	0.8570	0.8107	65.09	0.7965	0.6717
bs = 64†	**55.00**	**0.8004**	**0.5485**	9.2811	**82.84**	0.9076	**0.8313**	**78.05**	**0.8261**	**0.7871**	**83.04**	**0.8585**	**0.8148**	**65.09**	0.8021	**0.6717**
bs = 128	52.74	0.7954	0.5595	9.3429	82.84	0.9082	0.8313	74.68	0.8136	0.7587	82.94	0.8552	0.8148	65.09	0.7923	0.6717
bs = 192	52.74	0.7894	0.5595	**9.2598**	81.27	0.9076	0.8170	75.42	0.8146	0.7666	81.84	0.8509	0.8024	65.09	0.7951	0.6717
bs = 240	51.16	0.7891	0.5562	9.3096	81.27	**0.9093**	0.8170	74.00	0.8074	0.7526	82.94	0.8521	0.8148	65.09	**0.8028**	0.6717
(b) Pretext radiomic features																
Entropy	51.16	0.7936	0.5424	**9.1834**	78.39	0.9087	0.7814	73.46	0.8176	0.7461	80.00	0.8584	0.7900	63.75	0.7979	0.6430
Mean	51.35	0.7756	0.5147	9.3690	79.96	**0.9132**	0.8039	74.68	0.8116	0.7587	83.04	0.8549	0.8148	65.09	0.7979	0.6534
Var	52.74	0.7972	**0.5595**	9.2856	78.64	0.9104	0.7906	74.47	0.8139	0.7573	82.94	0.8546	0.8148	65.09	0.7979	0.6534
Mean + Var†	**55.00**	**0.8004**	0.5485	9.2811	**82.84**	0.9076	**0.8313**	**78.05**	**0.8261**	**0.7871**	**83.04**	**0.8585**	**0.8148**	**65.09**	**0.8021**	**0.6717**
(c) Brain atlas																
Desikan-Killiany	52.9	**0.8104**	**0.5629**	**9.1520**	**84.15**	0.8987	**0.8439**	72.99	0.8123	0.7414	80.94	0.8546	0.7983	**70.35**	0.7881	**0.7243**
Harvard-Oxford†	**55.00**	0.8004	0.5485	9.2811	82.84	**0.9076**	0.8313	**78.05**	**0.8261**	**0.7871**	**83.04**	**0.8585**	**0.8148**	65.09	**0.8021**	0.6717
(d) Covariance Estimation																
Population-level	52.93	0.7726	0.5345	**9.2172**	80.21	0.9076	0.8057	75.15	0.8181	0.7633	81.89	0.8534	0.8066	63.12	0.8017	0.6534
Minibatch-level†	**55.00**	**0.8004**	**0.5485**	9.2811	**82.84**	**0.9076**	**0.8313**	**78.05**	**0.8261**	**0.7871**	**83.04**	**0.8585**	**0.8148**	**65.09**	**0.8021**	**0.6717**

Transformers in medical MRI are increasingly used because their self-attention mechanism can capture both local tissue details and global anatomical context across entire 3D volumes. As such, following (Zhuang et al., 2025b), all methods use the same parameter initialization obtained from MAE pretraining for 300 epochs on the unlabeled pretraining dataset. All methods are pretrained on a single GPU, [NVIDIA RTX 6000 Ada Generation]

F. Ablation Study on Design Choices

We investigate which design choices are most responsible for the observed downstream gains of NAI-SSL. Specifically, we examine the roles of covariance estimation strategy, minibatch size, region-level radiomic descriptors, and neuroanatomical parcellation strategy by ablating four core components of the proposed objective: population- versus minibatch-level covariance estimation, covariance batch size, radiomic feature design, and atlas selection (Tab. 7). We first analyze the effect of minibatch size, which determines the stability of covariance estimation across samples. Performance remains stable across a broad range of batch sizes, while *a moderate batch size of 64 consistently provides the best trade-off between statistical stability and computational efficiency.* Next, for the choice of region-level radiomic descriptors, we find that simple intensity-based statistics consistently outperform more complex descriptors such as entropy. In particular, combining mean intensity and variance yields the strongest downstream performance. *These results suggest that low-order structural descriptors are sufficient to capture informative inter-hemispheric relationships.*

We further evaluate sensitivity to the underlying neuroanatomical atlas. Both the Harvard-Oxford (Rushmore et al., 2022) and Desikan-Killiany (Desikan et al., 2006) atlases achieve strong downstream performance, indicating that the proposed framework is robust across different anatomically defined parcellations. Overall, the Harvard-Oxford atlas performs better on most tasks, suggesting that anatomically meaningful parcellation can further improve the stability of covariance modeling. Then, we ablate the covariance estimation strategy by comparing minibatch-level covariance against population-level covariance matrix computed across the full pretraining dataset. Minibatch-level estimation consistently achieves better downstream performance. Overall, these ablations show that *moderate covariance-estimation batch sizes, simple mean–variance radiomic descriptors, and anatomically meaningful atlas selection are all useful for maximizing the effectiveness of NAI-SSL.*

Table 8. Ablation study on perturbation strategies. The default configuration (Masking) is marked with †. The best result in each column is **bolded**. BAcc = balanced accuracy (%). ↑: higher is better; ↓: lower is better.

Config	Gender			FTD		
	BAcc↑	AUC↑	F1↑	BAcc↑	AUC↑	F1↑
Noise	78.08	0.8847	0.7832	72.83	0.8087	0.7199
Masking†	**82.84**	**0.9076**	**0.8313**	**83.04**	**0.8585**	**0.8148**

Table 9. We repeat all downstream experiments across three independent runs on NAI-SSL. The mean performance is **bolded** and the Values in brackets represent the 95% confidence interval. BAcc = balanced accuracy (%). ↑: higher is better; ↓: lower is better.

	Age	Gender			MCI/CN		
Pretext	L1↓	BAcc↑	AUC↑	F1↑	BAcc↑	AUC↑	F1↑
NAI-SSL	**9.119** [0.421]	**82.837** [3.267]	**0.907** [0.001]	**0.831** [0.001]	**77.02** [3.86]	**0.817** [0.04]	**0.769** [0.04]

	FTD			Tumor Grade		
Pretext	BAcc↑	AUC↑	F1↑	BAcc↑	AUC↑	F1↑
NAI-SSL	**82.413** [2.848]	**0.847** [0.025]	**0.809** [0.024]	**63.56** [3.39]	**0.793** [0.023]	**0.656** [0.036]

	AD/MCI/CN (ext)			Age (ext)	MCI/CN (ext)		
Pretext	BAcc↑	AUC↑	F1↑	L1↓	BAcc↑	AUC↑	F1↑
NAI-SSL	**53.78** [2.94]	**0.797** [0.051]	**0.531** [0.015]	**9.545** [0.563]	**76.897** [0.843]	**0.864** [0.063]	**0.687** [0.067]

G. Ablation Study on Dataset Size

We analyze the robustness and data efficiency of NAI-SSL under varying amounts of downstream training data ranging from 40% to 100%. We then compare NAI-SSL against representative reconstruction- and contrastive-based SSL baselines (Fig. 3). *Across all data regimes, NAI-SSL achieves stronger or more stable performance than both MAE and SimCLR on downstream tasks.*

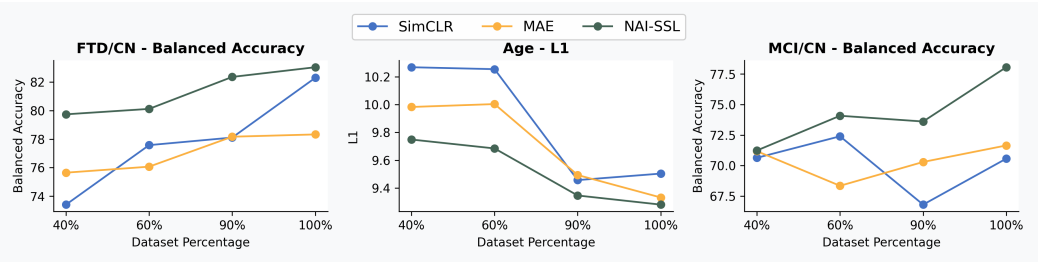

Figure 3. **Effect of downstream dataset size on representation transfer performance.** We evaluate SimCLR, MAE, and NAI-SSL under varying amounts of labeled downstream training data. Across neurodegenerative classification tasks and brain-age prediction, NAI-SSL consistently achieves stronger or more stable performance than standard SSL objectives across data regimes.

H. Inter-hemispheric Structural Covariance Analysis of Batch-wise Statistical Features

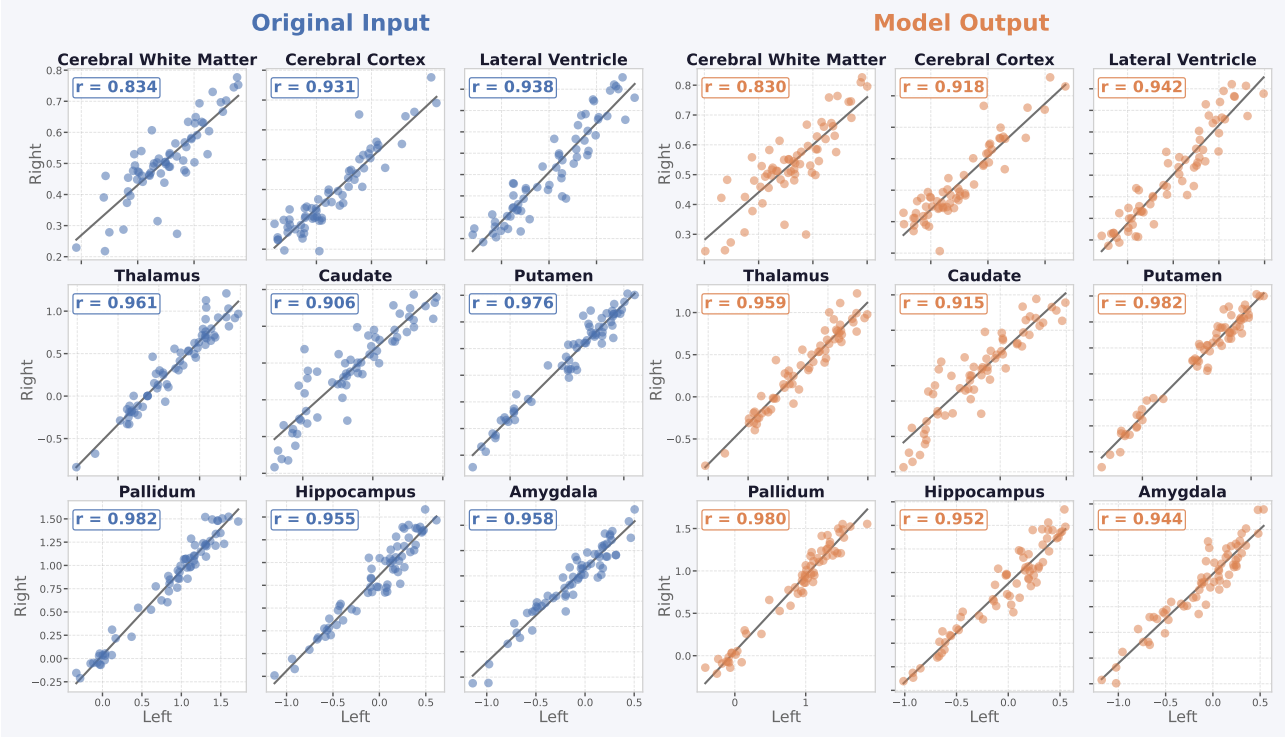

Figure 4. Inter-hemispheric correlation of ROI-level mean intensity features across nine atlas-defined brain regions within a representative minibatch. Left column (blue): original input data. Right column (orange): model output. Each subplot shows left- versus right-hemisphere ROI-level mean features across subjects, with the Pearson correlation coefficient r displayed in each panel. Best viewed in color and with zoom.

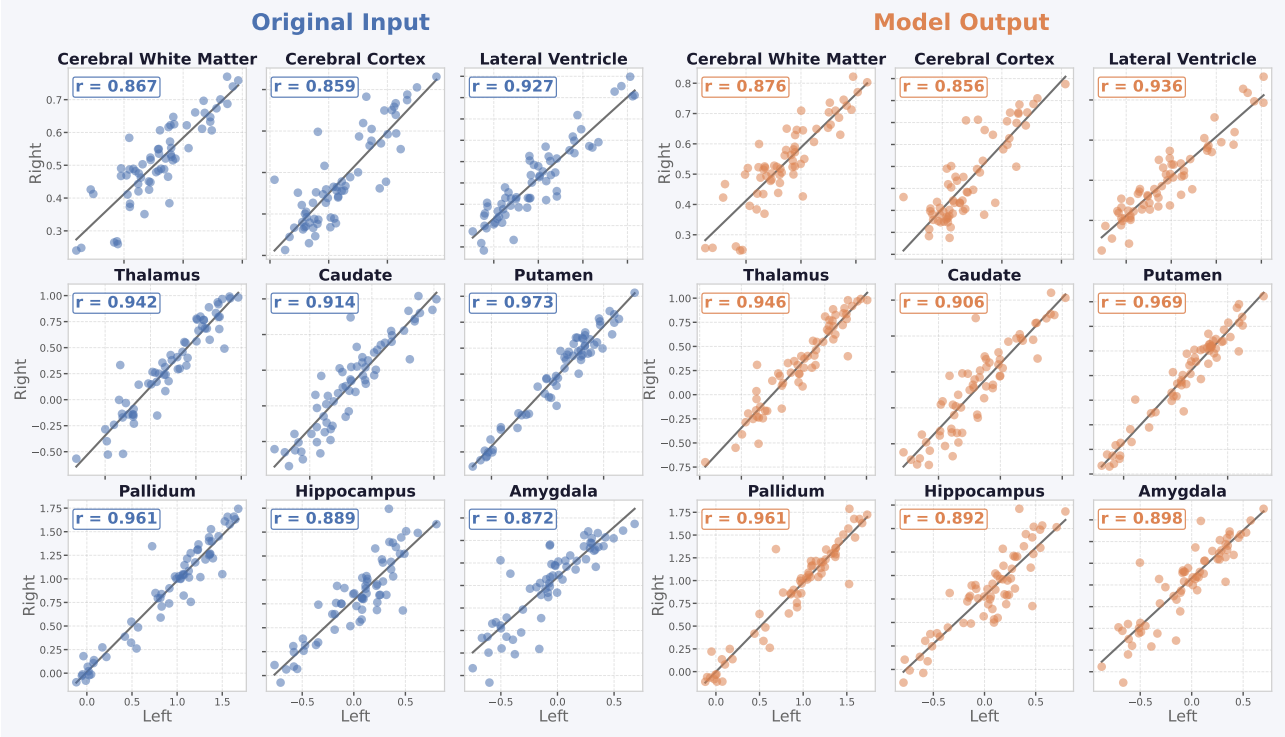

Figure 5. **Inter-hemispheric ROI mean-intensity correlation analysis within a representative minibatch.** Scatter plots show left- versus right-hemisphere ROI-level mean-intensity features across subjects for anatomically homologous brain regions. Pearson correlation coefficients (r) are shown for both the original input and model output, illustrating preservation of biologically meaningful inter-hemispheric relationships.

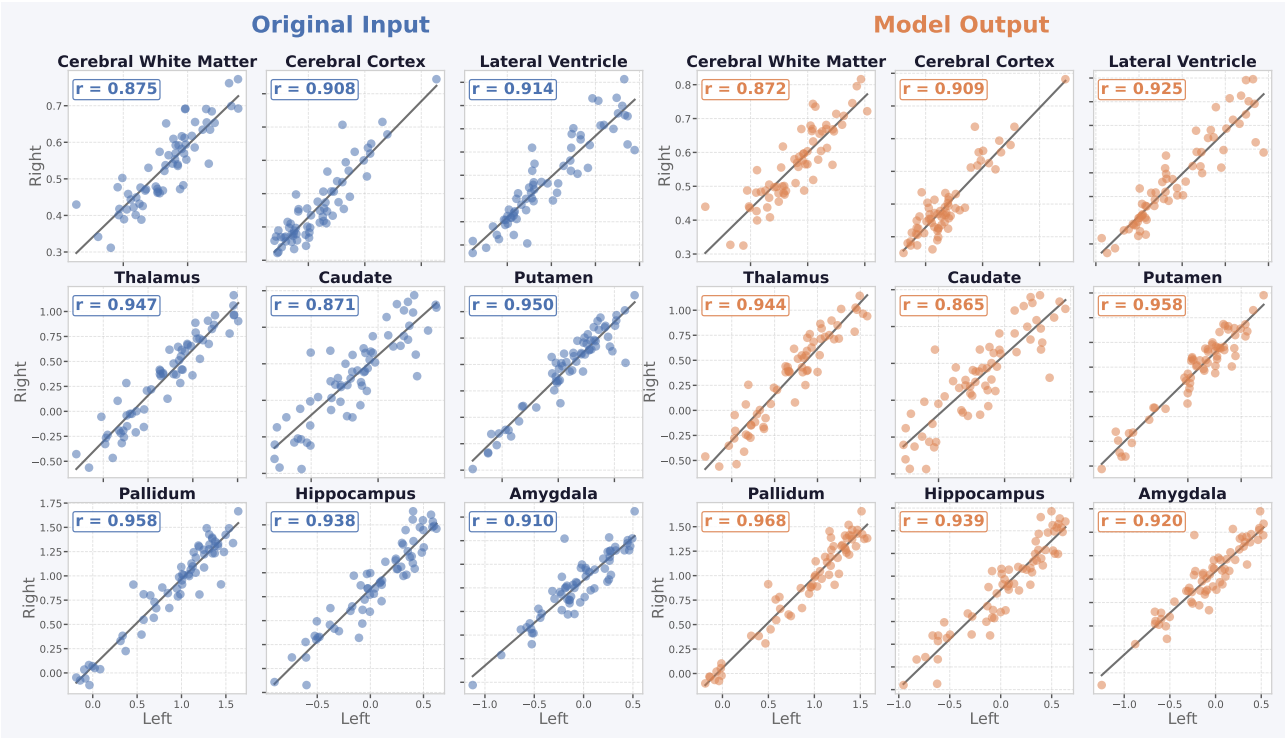

Figure 6. **Inter-hemispheric ROI mean-intensity correlation analysis within a representative minibatch.**

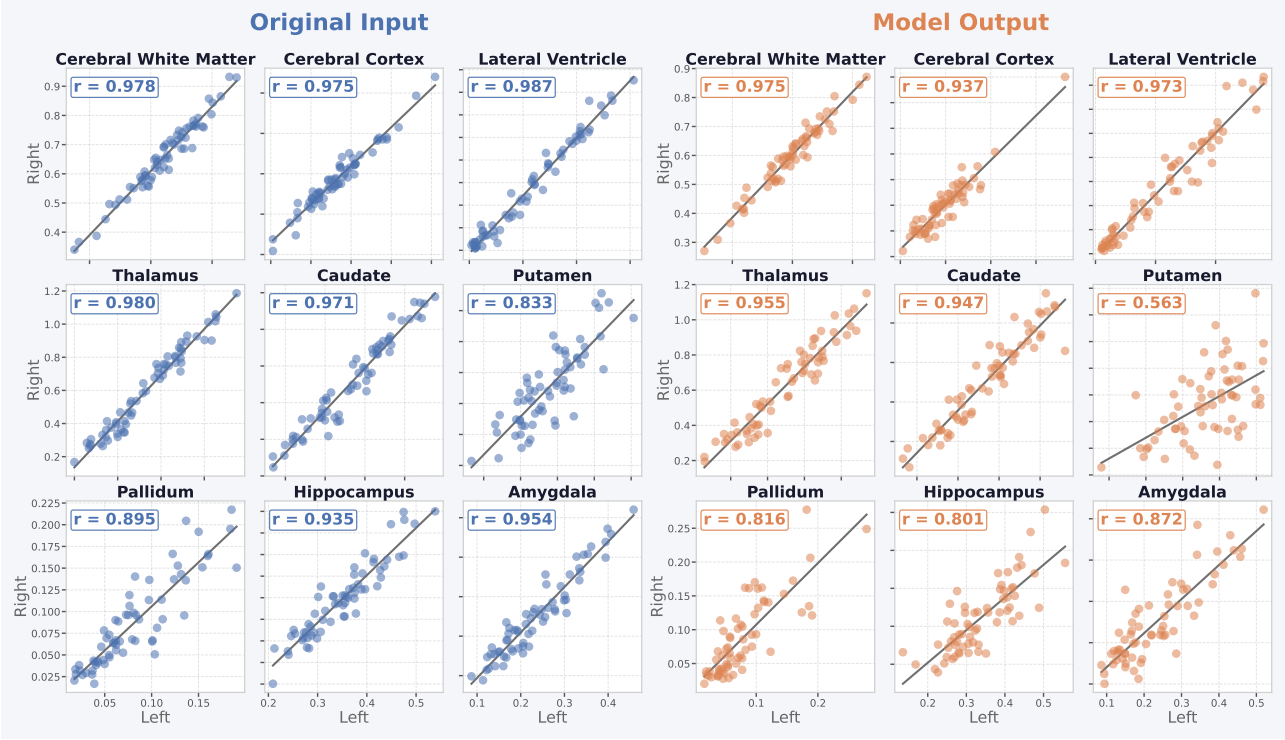

Figure 7. **Inter-hemispheric ROI variance correlation analysis within a representative minibatch.** Scatter plots show left- versus right-hemisphere ROI-level variance features across subjects for anatomically homologous brain regions. Pearson correlation coefficients (r) are shown for both the original input and model output, demonstrating preservation of inter-hemispheric structural covariance patterns.

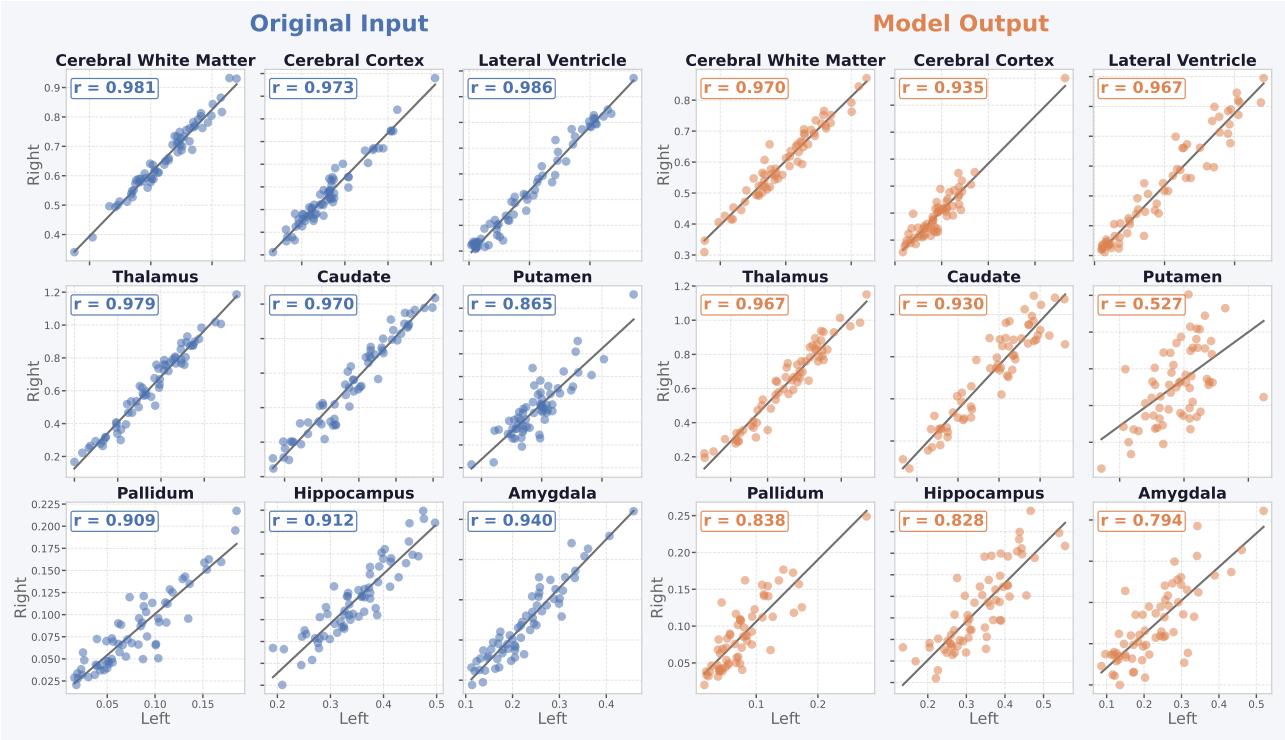

Figure 8. **Inter-hemispheric ROI variance correlation analysis within a representative minibatch.**

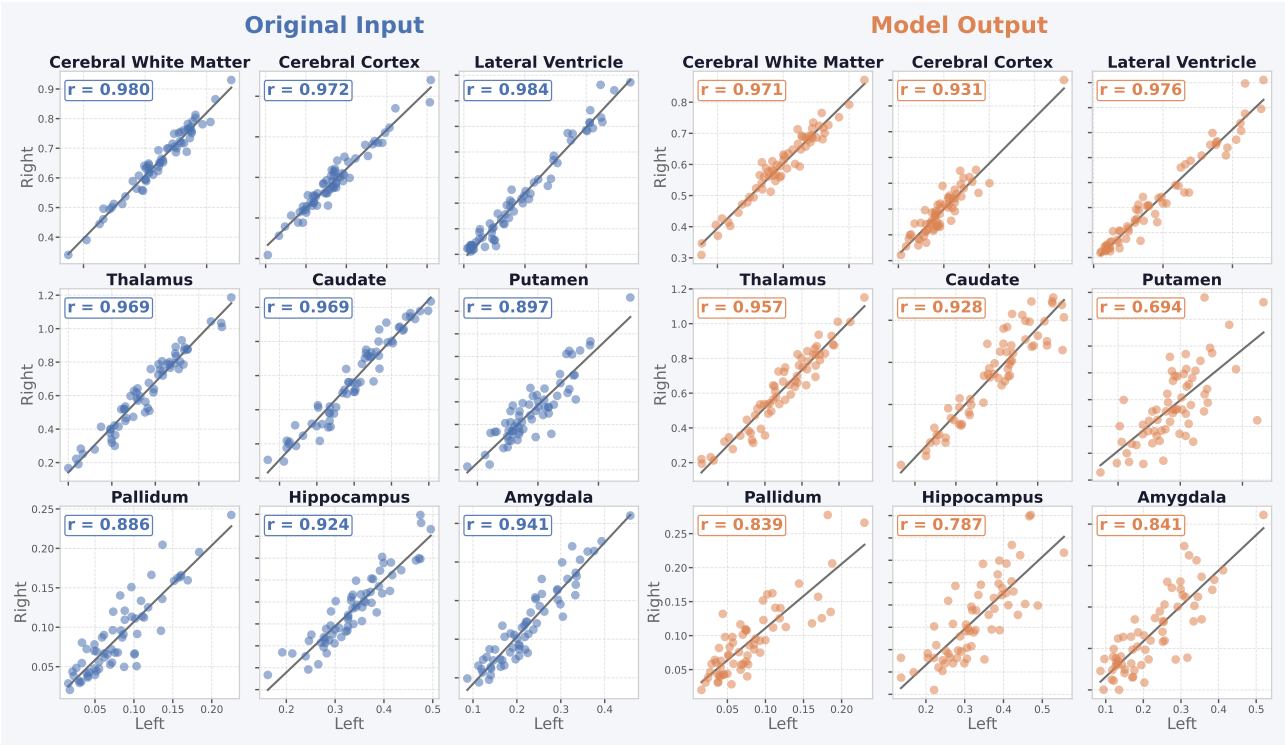

Figure 9. **Inter-hemispheric ROI variance correlation analysis within a representative minibatch.**

I. Limitations

The current implementation of NAI-SSL focuses primarily on inter-hemispheric structural consistency in structural brain MRI. While this provides a biologically grounded and effective supervisory signal, extending the framework to incorporate additional forms of neuroanatomical organization and other neuroimaging modalities remains an important direction for future work.

J. Impact Statement

This work uses only publicly available, fully de-identified structural MRI data and does not involve private or sensitive patient information. The goal of this research is to advance machine learning methods for neuroimaging analysis. While such methods may have downstream societal and clinical implications, we do not identify any immediate negative impacts arising directly from this work.