






PMIL: Prompt enhanced Multimodal Integrative analysis of fMRI combining functional connectivity and temporal Latency

Hyounghsin Choi^{1,2}, Jonghun Kim¹, Jiwon Chung¹, Bo-yong Park^{2,3},
and Hyunjin Park^{1,2*}

¹ Department of Electrical and Computer Engineering, Sungkyunkwan University, Suwon, South Korea

² Center for Neuroscience Imaging Research, Institute for Basic Science, Suwon, South Korea

³ Department of Brain and Cognitive Engineering, Korea University, Seoul, South Korea

{gudt1s17, hyunjinp}@skku.edu

Abstract. Functional connectivity (FC) analysis is the primary approach for studying functional magnetic resonance imaging (fMRI) data, focusing on the spatial patterns of brain activity. However, this method often neglects the temporal dynamics inherent in the timeseries nature of fMRI data, such as latency structure and intrinsic neural timescales (INT). These temporal features provide complementary insights into brain signals, capturing signal propagation and neural persistence information that FC alone cannot reveal. To address this limitation, we introduce Prompt enhanced multimodal integrative analysis (PMIL), a multimodal framework built on a transformer architecture that integrates latency structure and INT with conventional FC, enabling a more comprehensive analysis of fMRI data. Additionally, PMIL leverages text prompts within a state-of-the-art vision-language model to enhance the integration of INT with latency structure and FC. Our framework achieves state-of-the-art performance on an autism dataset, effectively distinguishing autistic patients from neurotypical individuals. Furthermore, PMIL identified disease-affected brain regions that align with findings from existing research, thereby enhancing its interpretability. The code for PMIL is publicly available at <https://github.com/gudt1s17/PMIL>.

Keywords: fMRI analysis · temporal dynamics · vision-language model · prompt tuning · autism spectrum disorder · functional connectivity.

1 Introduction

Functional magnetic resonance imaging (fMRI) measures blood oxygen level-dependent (BOLD) signals, serving as a surrogate for neural signals. As 4-D data comprising three spatial dimensions and one temporal dimension, fMRI

data is referred to as an fMRI timeseries. Previous studies have explored neural fluctuations through both spatial and temporal perspectives [13,15]. The primary analysis method for fMRI is functional connectivity (FC) analysis, which evaluates the synchronicity among brain regions to describe spatial patterns of brain activity [2,17,26]. While FC analysis has proven effective for numerous downstream tasks, existing approaches underrepresent the temporal dynamics of brain signals, though recent work increasingly addresses this aspect.

Temporal dynamics can be investigated through latency analysis, which measures delays or advances in the timeseries between brain regions [7]. This approach reveals the latency structure in brain signals and explains how brain signals propagate at a macroscopic level. By analyzing latency, it is possible to infer whether a signal in one region precedes a signal in another. In contrast, FC analysis assumes synchronicity and is effectively zero latency. Together, FC and latency structure analyses provide complementary insight into brain signals [22,24]. Another approach to examining temporal dynamics is the intrinsic neural timescale (INT), which measures the persistence of brain activity within a given region, typically assessed using autocorrelation [25]. It is widely used to evaluate functional specialization, as brain regions with longer timescales are better equipped for higher-order information like decision-making [29]. Unlike latency structure, INT reveals the intrinsic properties of individual regions rather than interregional interactions, making it a valuable complement to latency analysis.

Existing machine learning methods for fMRI analysis have predominantly focused on leveraging spatial patterns of FC for downstream tasks such as disease classification and clinical score regression. However, they often fail to fully utilize the complementary information provided by temporal dynamics [22,23,1,19,21,4]. In this study, we address this gap by integrating temporal aspects, including latency structure and INT, into fMRI analysis for the classification of autism spectrum disorder (ASD).

Recently, vision-language foundational models have led to significant advancements across various domains. These models have the potential to transform fMRI analysis as well. While existing foundational models trained on non-medical data are not directly applicable, medical vision-language models specifically designed for brain regions and their properties such as BiomedCLIP [31], offer a promising alternative [31,30,16]. In this work, we leverage BiomedCLIP by developing tailored text prompts to enhance the integration of temporal dynamics with the conventional FC. Our work addresses two key challenges:

1. The limited use of timeseries data, often neglecting temporal dynamics.
2. The sparse use of textual description in fMRI data analysis.

To overcome these limitations, we introduce Prompt enhanced multimodal integrative analysis (*PMIL*), a novel method for brain signal analysis built on a transformer framework. Transformers are well-suited for our multimodal approach, as they naturally integrate diverse features. Our key contributions are:

1. We extend brain signal analysis to incorporate temporal dynamics, including latency structure and INT, in addition to the spatial patterns of FC.

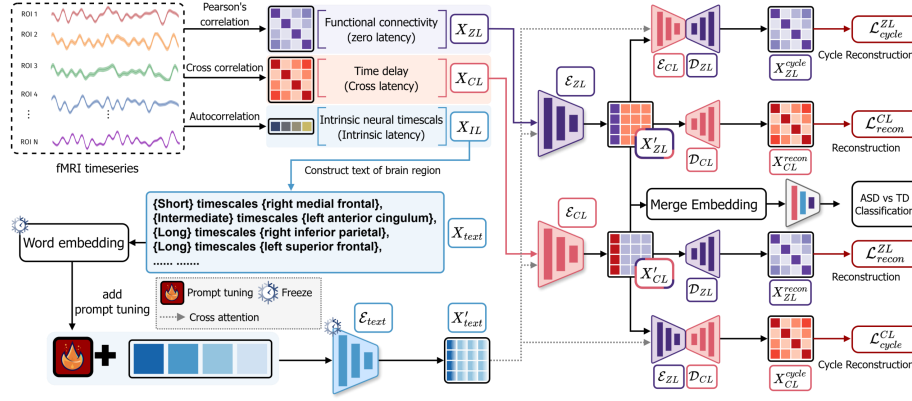


Fig. 1. Overview of the PMIL model. Derived from fMRI timeseries, zero and cross latency matrices are processed through their respective encoders. INT information is transformed into textual description, which undergoes prompt tuning. This text representation serves as a conditioning input for both encoders. The embedding from the zero latency encoder is used to reconstruct the original input (zero latency) and cross latency. Similarly, the embedding from the cross latency encoder is used to reconstruct the original input and zero latency. Finally, the embeddings from both encoders are merged and used for the downstream classification task. ASD: autism spectrum disorder, TD: typically developing group.

2. We design text prompts that effectively integrate INT with conventional FC, leveraging the capabilities of a recent vision-language model.
3. We demonstrate that our method outperforms existing baselines and highlights brain regions linked to ASD, thereby enhancing interpretability.

2 Method

2.1 Overview of the method

Using the pre-processed fMRI timeseries parcellated into N regions of interest (ROIs) based on a given atlas, we construct the FC (zero latency) matrix, $X_{ZL} \in R^{N \times N}$ by calculating Pearson correlation coefficients between ROI pairs and temporal delay (cross latency) matrix, $X_{CL} \in R^{N \times N}$ by maximizing the cross-correlation between different ROI pairs. Additionally, an INT (intrinsic latency) vector, $X_{IL} \in R^{N \times 1}$, is calculated using the autocorrelation for each ROI. X_{IL} is further used to generate a textual description $X_{text} \in R^N$. The matrices X_{ZL} and X_{CL} serve as inputs to encoders E_{ZL} and E_{CL} , respectively, with the text representation X'_{text} used as the conditioning input to generate embeddings within each encoder. The outputs from the encoders, X'_{ZL} and X'_{CL} are used to reconstruct the original cross latency matrix (X_{CL}^{recon}) and zero latency matrix (X_{ZL}^{recon}). For cycle-consistent reconstruction, embeddings X'_{ZL} are further used

in encoder E_{CL} to generate X'_{CL} , which is then used to reconstruct X'_{ZL} . Similarly, starting from X'_{CL} , the cycle reconstruction yields X'_{CL} . Finally, the embeddings from the zero and cross latency encoders are merged and passed through a pooling layer followed by multi-layer perceptrons (MLPs) to perform the classification task (see **Fig. 1**).

2.2 Computation of latency structure

Generating latency structure We measure the relationship between fMRI timeseries of two brain regions for both spatial and temporal perspectives. Given fMRI timeseries $\in R^{N \times t}$ where t represents the number of timepoints across N brain regions, the ij -th element of zero latency matrix X_{ZL} is given by

$$r_{i,j} = \text{pearson correlation}(\text{timeseries}_i, \text{timeseries}_j), \quad i, j \in 1, 2, \dots, N \quad (1)$$

Following established protocols for investigating latency structure [22,24], the ij -th element of cross latency matrix X_{CL} is given by

$$C_{i,j}(\tau) = \arg\max \frac{1}{T} \int \text{timeseries}_i(t + \tau) \cdot \text{timeseries}_j(t) dt, \quad i, j \in 1, 2, \dots, N \quad (2)$$

We used τ as the element of X_{CL} that maximized $C_{i,j}(\tau)$, where T is the length of the fMRI scan. Each element of X_{IL} is computed using the autocorrelation of a given brain region by making the region indices i and j the same in eq (2). It is computed as τ when $C_{i,j}(\tau)$ is approximately 0. In other words, X_{IL} represents the rate at which the autocorrelation of a brain region decays.

Generating text description To utilize the rich information in pretrained vision-language models, we generate a text description X_{text} from the intrinsic latency vector X_{IL} . Since the range of values carries more significance than specific values and can be effectively represented as a low-dimensional vector, we design practical text descriptions. To emphasize regions with strong FC, we retained the top 5% of connections in the zero latency matrix and further selected regions with the top 5% row-wise sum values [3,14,27]. These regions were then used to generate the text descriptions like "<short> timescales <right medial frontal region>". Consistent with prior research [3,14,27], we categorized intrinsic latency values shorter than 3 seconds as 'short', between 3 and 5 seconds as 'intermediate', and longer than 5 seconds as 'long'.

2.3 Encoders, decoders, and embeddings

Encoder for latency representation A standard transformer decoder layer is used to generate embeddings (X'_{ZL} and X'_{CL}) for the two input types [28]. The zero latency matrix serve as the input to encoder E_{ZL} , while the cross latency matrix is used as the input to encoder E_{CL} , along with text representation X'_{text} . The output of BiomedCLIP pretrained word embedding tokenizer and

text encoder are used to generate text representation X'_{text} . This representation facilitates both E_{ZL} and E_{CL} by providing additional contextual information for k -selected regions associated with INT. To enhance the quality of the text representation, prompt tuning is applied, incorporating a prompt token p .

$$X'_{text} = E_{text}([p, \text{Word embedding}(X_{text})]), \quad X'_{text} \in R^{k \times N} \quad (3)$$

$$X'_{ZL} = E_{ZL}(X_{ZL}, X'_{text}), \quad X'_{CL} = E_{CL}(X_{CL}, X'_{text}) \quad (4)$$

Decoder layer for reconstruction A standard transformer decoder layer was used for the decoder layer. Since the zero latency matrix and cross latency matrix are physically related [22,24], we use decoders D_{CL} and D_{ZL} to reconstruct one matrix from the embedding of the other, thereby strengthening their interactions (eq. 5). However, if the embeddings are used solely for single-directional reconstruction, the features may become biased. To address this, embeddings from one type are encouraged to generate embeddings of the other type, followed by a reconstruction back to the original input type, completing a cycle-consistent reconstruction (eq. 6) [32]. That is X'_{ZL} is passed through to E_{CL} and subsequently D_{ZL} to reconstruct X_{ZL}^{cycle} . The same applies to X'_{CL} .

$$X_{CL}^{recon} = D_{CL}(X'_{ZL}), \quad X_{ZL}^{recon} = D_{ZL}(X'_{CL}) \quad (5)$$

$$X_{ZL}^{cycle} = D_{ZL}(E_{CL}(X'_{ZL}, X'_{text})), \quad X_{CL}^{cycle} = D_{CL}(E_{ZL}(X'_{CL}, X'_{text})) \quad (6)$$

Merging embeddings and readout layer We adopt a learnable weighted sum to determine the optimal proportion for merging two embeddings (eq. 7). Both latency types are structured as region-by-region graphs. To process these, we apply the OCRead layer, which is specifically designed for graph-based data by soft-clustering token embeddings and projecting them orthogonally to produce meaningful graph-level embeddings. This method outperforms class token approaches [1,19]. The resulting embeddings are then passed through MLPs for classification.

$$X_{merge} = \lambda X'_{ZL} + (1 - \lambda) X'_{CL} \quad (7)$$

2.4 Loss function

We employ tailored loss functions for different model objectives. Mean squared error (MSE) for reconstruction (L_{recon}), L1 loss for cycle consistency reconstruction (L_{cycle}), and cross-entropy loss for the classification prediction (L_{pred}). The total loss function is defined as $L = \lambda_1 L_{recon} + \lambda_2 L_{cycle} + \lambda_3 L_{pred}$.

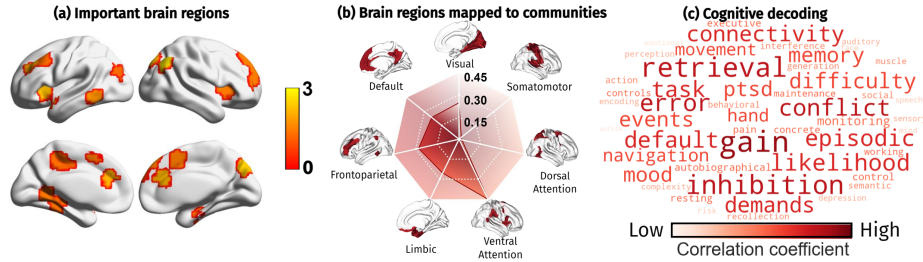
3 Experiments

3.1 Datasets and experimental settings

Dataset We evaluated our method using the publicly available fMRI dataset on ASD, Autism Brain Imaging Data Exchange (ABIDE) [9]. This dataset aggregates data from 17 international sites and has been preprocessed using the

Table 1. Quantitative diagnosis prediction results of our method compared with baseline models (Mean \pm standard deviation). V&L: Vision and Language.

Model	Category	Accuracy	AUROC	Sensitivity	Specificity
BrainNetCNN(NImg17) [20]	Vision	66.6 \pm 2.1	72.0 \pm 3.0	72.9 \pm 7.0	59.3 \pm 10.0
FBNETGNN(MIDL22) [18]	Vision	66.8 \pm 1.3	73.5 \pm 0.7	68.3 \pm 5.0	65.1 \pm 5.2
BNT(NeurIPS22) [19]	Vision	73.4 \pm 2.0	81.5 \pm 1.3	79.8 \pm 6.2	66.0 \pm 11.2
Com-BrainTF(MICCAI23) [1]	Vision	72.1 \pm 3.4	81.1 \pm 2.5	79.8\pm3.1	64.0 \pm 6.9
Ours	V&L	75.9\pm1.9	83.4\pm0.5	79.6 \pm 3.7	71.2\pm3.4

**Fig. 2.** The interpretation of the ROI-level importance. (a) illustrates the top 10% important regions in predicting ASD. (b) illustrates brain regions assigned to brain functional communities. (c) demonstrated the cognitive decoding of brain regions.

Configurable Pipeline for the Analysis of Connectomes (CPAC) software [7]. Brain regions were parcellated using the Craddock 200 atlas [8]. We use a total of 1009 participants, of whom 516 (51%) were diagnosed with ASD. Given the multi-site nature of the dataset, with data collected using various scanners and acquisition parameters, we implemented techniques to mitigate site variability. First, covariate control was applied to adjust for site effects, as well as participant age and sex, across FC (zero latency), temporal delay (cross latency), and INT (intrinsic latency) [12]. Second, a stratified sampling strategy was implemented during the training-validation-test split to maintain the ratio of ASD to typically developing participants across different collection sites [1,19].

Implementation Details All models were implemented in PyTorch and trained using NVIDIA RTX 4070 TI (12GB). Each transformer layer was configured with 4 attention heads. The Adam optimizer was used, with an initial learning rate of 10^{-4} and a weight decay of 10^{-4} . The batch size was set to 32, and the dataset was split into training, validation, and test sets in a ratio of 7: 1: 2. We empirically set λ_1 and λ_2 to 0.1 and λ_3 to 1. Training was conducted over 50 epochs and the model achieving the highest area under the receiver operator curve (AUROC) performance on the validation set was selected for comparison on the test set. We reported an average of 5 random runs on the test set.

Table 2. Ablation studies on the combination of different input types.
 IL: intrinsic latency, CL: cross latency, ZL: zero latency. V&L: Vision and Language.

Input types	Category	Accuracy	AUROC	Sensitivity	Specificity
IL	Language	52.1±2.6	48.3±3.0	60±49.0	40±49.0
CL	Vision	53.1±3.2	55.4±1.5	44.7±23.0	62.9±19.4
ZL	Vision	70.6±3.7	81.4±1.4	83.0±7.2	57.4±12.5
CL+IL	V&L	53.9±2.3	54.9±1.7	59.6±5.2	47.5±4.7
ZL+CL	Vision	74.0±2.8	82.2±1.8	81.0±5.5	65.7±9.9
ZL+IL	V&L	75.0±1.6	82.8±0.5	78.5±3.6	71.0±3.1
Ours (ZL+CL+IL)	V&L	75.9±1.9	83.4±0.5	79.6±3.7	71.2±3.4

3.2 Experimental results

Comparison with State-of-the-art Methods The quantitative diagnosis prediction results are presented in **Table 1**. Our model achieved the best performance in terms of accuracy, AUROC, and specificity. While improvements in sensitivity were marginal, our approach outperformed other baselines, even those using the same transformer architecture. This superior performance is attributed to our integration of temporal dynamics alongside the conventional FC and the inclusion of intrinsic latency text descriptions generated with a pretrained language model [1,19]. These results emphasize the added value of leveraging temporal dynamics in fMRI analysis.

Interpretation of the ROI-level importance To assess the importance of ROI in predicting ASD, we visualized the important regions using the merged embeddings generated through the class token. **Fig. 2(a)** visualizes the top 10% important regions. The angular gyrus, anterior and medial cingulate, superior, and medial frontal gyrus, superior and medial temporal gyrus, lingual gyrus, cuneus, insula, fusiform, supplementary motor area, and medial occipital gyrus were emphasized. These emphasized regions have been previously identified as related to ASD diagnosis and clinical symptoms [10,11,6,5]. **Fig. 2(b)** illustrates important regions stratified across seven brain functional communities, with the ventral attention network receiving the greatest emphasis. This observation aligns with existing research, which identifies atypical development in this network as a hallmark characteristic of ASD [10,11]. **Fig. 2(c)** demonstrated the cognitive decoding of important regions, indicating that the highlighted areas are associated with cognitive functions often impaired in ASD, such as emotional inhibition, executive functioning, and memory retrieval [6].

Ablation study We performed ablation studies to examine the effects of data types, reconstruction, text description, and embedding merge strategies. **Table 2** highlights performance variations based on input data types showing that using cross or intrinsic latency individually—or even in combination—does not achieve strong predictive performance. However, these input types enhance prediction when integrated with zero latency (i.e., FC). In summary, temporal features

Table 3. Ablation studies about loss, text description, and merging embeddings.

Method	Accuracy	AUROC	Sensitivity	Specificity
W/O cycle loss	74.8 \pm 2.3	83.2 \pm 1.1	80.6 \pm 5.9	68.2 \pm 7.1
W/O recon and cycle loss	72.4 \pm 3.3	80.6 \pm 2.9	76.1 \pm 6.9	68.6 \pm 8.8
W/O text (real valued ZL)	73.6 \pm 2.5	81.4 \pm 1.6	81.5\pm3.7	64.5 \pm 8.5
Merging embeddings	Accuracy	AUROC	Sensitivity	Specificity
Merging by mean	72.1 \pm 0.9	80.2 \pm 0.7	79.3 \pm 4.0	64.0 \pm 5.9
Merging by concatenate	68.6 \pm 2.1	78.8 \pm 1.7	72.9 \pm 9.1	64.2 \pm 10.1
Merging by sum	70.4 \pm 2.1	79.5 \pm 1.6	77.0 \pm 6.2	62.5 \pm 8.3
Merging by max	71.3 \pm 2.1	77.9 \pm 2.6	72.0 \pm 5.1	70.2 \pm 1.9
Ours	75.9\pm1.9	83.4\pm0.5	79.6 \pm 3.7	71.2\pm3.4

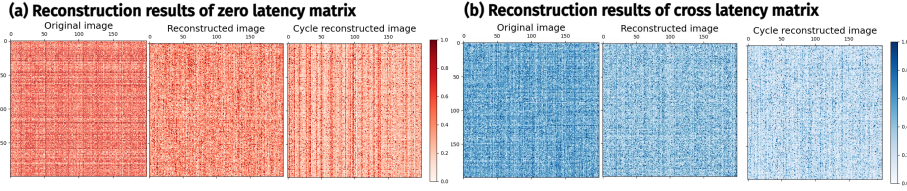


Fig. 3. The results of zero latency and cross latency matrix reconstruction, (a) illustrates the reconstruction results of zero latency matrix. Reconstructed image is the result of X'_{CL} fed to D_{ZL} and cycle reconstructed image is the result of X'_{ZL} fed to E_{CL} followed by D_{ZL} . (b) illustrates the reconstruction results of cross latency matrix. Reconstructed image is the result of X'_{ZL} fed to D_{CL} and cycle reconstructed image is the result of X'_{CL} fed to E_{ZL} followed by D_{CL} .

and text description enhance accuracy, AUROC, and specificity. **Table 3** evaluates the impact of different combinations of loss terms, including the use of text description and embedding merge approaches. Removing reconstruction or cycle-consistent reconstruction losses results in reduced performance, with reconstruction loss having a more considerable impact on accuracy and AUROC, while cycle-consistent loss better balances sensitivity and specificity. **Fig. 3** visualizes the result of reconstruction and cycle reconstruction. The MSE between original and reconstructed data are 0.042 ± 0.018 for zero latency reconstruction, 0.039 ± 0.019 for cycle-consistent reconstruction, and 0.086 ± 0.016 for cross latency, indicating small reconstruction errors. Moreover, incorporating text descriptions of intrinsic latency boosts performance, emphasizing the potential of using a pretrained text encoder for further refinement in disorder prediction. Lastly, the learnable weighted sum method for merging embeddings outperforms other methods, demonstrating the effectiveness of the parametrized approach.

4 Conclusion

This study presents a novel fMRI analysis method integrating the temporal dynamics (cross and intrinsic latency) with the spatial FC patterns. PMIL demonstrated superior performance in predicting ASD compared to typically develop-

ing individuals while identifying explainable and clinically relevant brain regions. Moreover, our method linked these regions to cognitive functions commonly associated with ASD. A key feature of PMIL is using intrinsic latency-based text descriptions via a pretrained text encoder to enhance multi-modal brain analysis. This highlights the potential of PMIL for application to a wide range of psychiatric disorders beyond ASD. Future work will focus on improving the generalizability, exploring diverse text description strategies, and incorporating different brain atlases to expand its utility across various neuroimaging datasets.

Acknowledgments. This study was supported by National Research Foundation (RS-2024-00408040), AI Graduate School Support Program (Sungkyunkwan University) (RS-2019-II190421), ICT Creative Consilience program (RS-2020-II201821), and Artificial Intelligence Innovation Hub program (RS-2021-II212068).

Disclosure of Interests. The authors declare no competing interests relevant to the content of this article.

References

1. Bannadabhavi, A., Lee, S., Deng, W., Ying, R., Li, X.: Community-aware transformer for autism prediction in fmri connectome. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 287–297. Springer (2023)
2. Breakspear, M.: Dynamic models of large-scale brain activity. *Nature neuroscience* **20**(3), 340–352 (2017)
3. Cavanagh, S.E., Hunt, L.T., Kennerley, S.W.: A diversity of intrinsic timescales underlie neural computations. *Frontiers in Neural Circuits* **14**, 615626 (2020)
4. Choi, H., Byeon, K., Lee, J.e., Hong, S.J., Park, B.y., Park, H.: Identifying subgroups of eating behavior traits unrelated to obesity using functional connectivity and feature representation learning. *Human Brain Mapping* **45**(1), e26581 (2024)
5. Choi, H., Byeon, K., Park, B.y., Lee, J.e., Valk, S.L., Bernhardt, B., Di Martino, A., Milham, M., Hong, S.J., Park, H.: Diagnosis-informed connectivity subtyping discovers subgroups of autism with reproducible symptom profiles. *NeuroImage* **256**, 119212 (2022)
6. Cooper, R.A., Richter, F.R., Bays, P.M., Plaisted-Grant, K.C., Baron-Cohen, S., Simons, J.S.: Reduced hippocampal functional connectivity during episodic memory retrieval in autism. *Cerebral Cortex* **27**(2), 888–902 (2017)
7. Craddock, C., Benhajali, Y., Chu, C., Chouinard, F., Evans, A., Jakab, A., Khundrakpam, B.S., Lewis, J.D., Li, Q., Milham, M.: The neuro bureau preprocessing initiative: open sharing of preprocessed neuroimaging data and derivatives. *Frontiers in Neuroinformatics* **7**(27), 5 (2013)
8. Craddock, R.C., James, G.A., Holtzheimer III, P.E., Hu, X.P., Mayberg, H.S.: A whole brain fmri atlas generated via spatially constrained spectral clustering. *Human brain mapping* **33**(8), 1914–1928 (2012)
9. Di Martino, A., Yan, C.G., Li, Q., Denio, E., Castellanos, F.X., Alaerts, K., Anderson, J.S., Assaf, M., Bookheimer, S.Y., Dapretto, M.: The autism brain imaging data exchange: towards a large-scale evaluation of the intrinsic brain architecture in autism. *Molecular psychiatry* **19**(6), 659–667 (2014)

10. Doyle-Thomas, K.A., Lee, W., Foster, N.E., Tryfon, A., Ouimet, T., Hyde, K.L., Evans, A.C., Lewis, J., Zwaigenbaum, L., Anagnostou, E.: Atypical functional brain connectivity during rest in autism spectrum disorders. *Annals of neurology* **77**(5), 866–876 (2015)
11. Farrant, K., Uddin, L.Q.: Atypical developmental of dorsal and ventral attention networks in autism. *Developmental science* **19**(4), 550–563 (2016)
12. Fortin, J.P., Cullen, N., Sheline, Y.I., Taylor, W.D., Aselcioglu, I., Cook, P.A., Adams, P., Cooper, C., Fava, M., McGrath, P.J.: Harmonization of cortical thickness measurements across scanners and sites. *Neuroimage* **167**, 104–120 (2018)
13. Fox, M.D., Raichle, M.E.: Spontaneous fluctuations in brain activity observed with functional magnetic resonance imaging. *Nature reviews neuroscience* **8**(9), 700–711 (2007)
14. Golesorkhi, M., Gomez-Pilar, J., Zilio, F., Berberian, N., Wolff, A., Yagoub, M.C., Northoff, G.: The brain and its time: intrinsic neural timescales are key for input processing. *Communications biology* **4**(1), 970 (2021)
15. Greicius, M.D., Krasnow, B., Reiss, A.L., Menon, V.: Functional connectivity in the resting brain: a network analysis of the default mode hypothesis. *Proceedings of the national academy of sciences* **100**(1), 253–258 (2003)
16. He, Z., Li, W., Liu, Y., Liu, X., Han, J., Zhang, T., Yuan, Y.: Fm-app: Foundation model for any phenotype prediction via fmri to smri knowledge transfer. *IEEE Transactions on Medical Imaging* (2024)
17. Honey, C.J., Kötter, R., Breakspear, M., Sporns, O.: Network structure of cerebral cortex shapes functional connectivity on multiple time scales. *Proceedings of the National Academy of Sciences* **104**(24), 10240–10245 (2007)
18. Kan, X., Cui, H., Lukemire, J., Guo, Y., Yang, C.: Fbnetgen: Task-aware gnn-based fmri analysis via functional brain network generation. In: *International Conference on Medical Imaging with Deep Learning*. pp. 618–637. PMLR (2022)
19. Kan, X., Dai, W., Cui, H., Zhang, Z., Guo, Y., Yang, C.: Brain network transformer. *Advances in Neural Information Processing Systems* **35**, 25586–25599 (2022)
20. Kawahara, J., Brown, C.J., Miller, S.P., Booth, B.G., Chau, V., Grunau, R.E., Zwicker, J.G., Hamarneh, G.: Brainnetcnn: Convolutional neural networks for brain networks; towards predicting neurodevelopment. *NeuroImage* **146**, 1038–1049 (2017)
21. Liu, R., Huang, Z.A., Hu, Y., Huang, L., Wong, K.C., Tan, K.C.: Spatio-temporal hybrid attentive graph network for diagnosis of mental disorders on fmri time-series data. *IEEE Transactions on Emerging Topics in Computational Intelligence* (2024)
22. Mitra, A., Snyder, A.Z., Blazey, T., Raichle, M.E.: Lag threads organize the brain’s intrinsic activity. *Proceedings of the National Academy of Sciences* **112**(17), E2235–E2244 (2015)
23. Mitra, A., Snyder, A.Z., Constantino, J.N., Raichle, M.E.: The lag structure of intrinsic activity is focally altered in high functioning adults with autism. *Cerebral cortex* **27**(2), bhv294 (2015)
24. Mitra, A., Snyder, A.Z., Hacker, C.D., Raichle, M.E.: Lag structure in resting-state fmri. *Journal of neurophysiology* **111**(11), 2374–2391 (2014)
25. Murray, J.D., Bernacchia, A., Freedman, D.J., Romo, R., Wallis, J.D., Cai, X., Padoa-Schioppa, C., Pasternak, T., Seo, H., Lee, D.: A hierarchy of intrinsic timescales across primate cortex. *Nature neuroscience* **17**(12), 1661–1663 (2014)
26. Smith, S.M., Fox, P.T., Miller, K.L., Glahn, D.C., Fox, P.M., Mackay, C.E., Filippini, N., Watkins, K.E., Toro, R., Laird, A.R.: Correspondence of the brain’s func-

- tional architecture during activation and rest. *Proceedings of the national academy of sciences* **106**(31), 13040–13045 (2009)
27. Soltani, A., Murray, J.D., Seo, H., Lee, D.: Timescales of cognition in the brain. *Current opinion in behavioral sciences* **41**, 30–37 (2021)
28. Vaswani, A.: Attention is all you need. *Advances in Neural Information Processing Systems* (2017)
29. Wang, X.J.: Probabilistic decision making by slow reverberation in cortical circuits. *Neuron* **36**(5), 955–968 (2002)
30. Wang, Z., Wu, Z., Agarwal, D., Sun, J.: Medclip: Contrastive learning from unpaired medical images and text. In: *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. pp. 3876–3887 (2022)
31. Zhang, S., Xu, Y., Usuyama, N., Xu, H., Bagga, J., Tinn, R., Preston, S., Rao, R., Wei, M., Valluri, N.: Biomedclip: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. *arXiv preprint arXiv:2303.00915* (2023)
32. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: *Proceedings of the IEEE international conference on computer vision*. pp. 2223–2232 (2017)