Neuro Anatomical Transformers

Paul Fingl [1 2] **Matei Stan** [1 3] **Leon Luithlen** [1] **Catalin Mitelut** [1]

Abstract

Understanding large-scale brain dynamics requires models that capture nonlinear structure while remaining interpretable and biologically grounded. Transformer architectures have shown promise for time series modeling but offer limited interpretability, restricting their use in causal neuroscience and AI safety. We present Neuro-Anatomical Transformers (NATs), a modular architecture designed to model pairwise brain-region interactions while preserving transparency. NATs use Region-Attention Blocks (RABs), which mirror asymmetric neuroanatomy and assign dedicated attention heads to distinct region–region pathways. This structure maintains separable, traceable information flow across layers, enabling direct analysis of internal circuits. Unlike traditional AI models, which often function as opaque black boxes, NATs support white box analysis and perturbation which is key for identifying misalignment, reward hacking, or failure modes. These tools are essential for developing verifiable, interpretable AI systems, a growing need in AI safety. Current interpretability approaches often retrofit transparency onto uninterpretable models; by contrast, NATs are transparent by design. We view this work as a technical stepping stone toward whole-brain emulation, which may offer a reference architecture for safer agentic AI systems. NATs thus advance both neuroscience and AI alignment by linking biological fidelity with structural transparency.

1. Introduction

Understanding large-scale, brain-wide neural dynamics requires both high-resolution experimental data and computational tools capable of modeling the highly nonlinear and distributed properties of neural systems. While transformer architectures have demonstrated remarkable success in capturing long-range dependencies in domains such as natural language (Vaswani et al., 2017a), their application to neuroscience remains limited.

Recent efforts in the burgeoning field of *NeuroAI*, which bridges neural data and artificial intelligence, have begun to pave the way with transformer-based approaches to neural representations. For example, the Neural Data Transformer 2 (NDT2) specializes in modeling spatiotemporal structure in neural spiking data and excels in transfer across sessions, subjects, and behavioral contexts (Ye & Pandarinath, 2023). Similarly, SwiFT, a Swin-style 4D fMRI transformer, directly learns from volumetric brain activity to predict behavioral traits while maintaining spatial interpretability (Kim et al., 2023). Transformer models have also been applied to brain network graphs—such as in the Brain Network Transformer, which uses pairwise connectivities between regions of interest (ROIs) to perform attention computations and improve classification tasks (Kan et al., 2022). Several studies have also demonstrated how enhancing transformer models with sparse autoencoders applied to calcium imaging data can yield more interpretable latent features aligned with neuroscience variables like stimulus orientation (Freeman et al., 2025). At a broader theoretical level, the emerging field of mechanistic interpretability in transformers is drawing deeply from dynamical systems theory, positioning transformer residual streams as evolving trajectories (akin to neural dynamics) to uncover mechanisms underlying model behavior (Fernando & Guitchounts, 2025).

There are several key challenges that remain including: (*i*) the lack of architectures explicitly aligned with neuroanatomical connectivity, and (*ii*) the difficulty of interpreting learned representations in ways that support causal, perturbation-driven analyses.

Here we focus on these specific challenges, in particular, on investigating transformer-based NN architectures that mimic or better capture the asymmetric bi-directional interactions between brain regions. We introduce the **Neuro-Anatomical Transformer (NAT)**, a transformer-based neural network explicitly designed to model interactions between distinct brain regions. Central to NATs are the **Region-Attention-Block (RAB)**, a modular attention mechanism that models all pairwise region interactions via a

[1] Netholabs, London, UK [2] Technical University of Munich, Germany [3] University of Manchester, UK. Correspondence to: Catalin Mitelut <cat@netholabs.com>.

type of cross-model attention. Each RAB is designed to learn and capture distinct region-region pair interactions, effectively mimicking the asymmetric interactions observed across brain areas. This design ensures that distinct information channels are preserved deep into the network, yielding interpretable components. Importantly, this architecture provides a natural pathway for *in silico* perturbations, enabling causal hypothesis testing on modeled dynamics, and can be generalized to multimodal inputs for building more comprehensive "world models" of neural computation.

We describe the NAT architecture below (see Section 2). To demonstrate the potential of our approach, we applied NATs to widefield calcium imaging data across 16 mouse brain regions (Mitelut et al., 2022) (see Section 3). We find that NATs match or can improve the performance of standard transformers while offering significantly enhanced interpretability and neuroanatomical fidelity.

1.1. NATs for AI Safety

We view NATs not only as a novel tool for modeling mesoscale brain dynamics, but also as a technical contribution to the broader goals of AI safety and interpretability. Their modular, anatomically grounded architecture makes them a natural fit for building white-box cognitive models, where internal processes can be traced, perturbed, and understood. This interpretability is crucial for identifying and analyzing alignment-relevant failure modes such as reward misgeneralization, deceptive behavior, and goal drift, which are difficult to study in standard black-box deep learning systems. Unlike conventional transformers that entangle computations across layers, NATs preserve separable information channels. This allows for targeted circuit-level interventions and causal analysis, which are essential features for building auditable agentic systems aligned with human intentions.

1.2. NATs for whole-brain-emulation

We also view NATs as a potential building block for whole-brain emulation (WBE). Their biologically informed, region-level design mirrors the structure of real nervous systems and provides a scalable foundation for simulating increasingly complex brains in silico. NATs can integrate real neural recordings into predictive models while preserving anatomical and functional structure, paving the way for cognitive emulators with known internal dynamics. As WBE emerges as a potential alternative path to AGI, NATs offer a transparent and structured approach to modeling brain-like agents. These models could serve as safety testbeds and reference architectures for agentic AI systems whose goals and reasoning can be understood, influenced, and aligned with human values.

1.3. NATs for interpretability

We propose that NAT design could offer some advantages over standard transformer multi-head-attention (MHA) block desigsn.

Distribution of computational load. The NAT design distributes the computational load while maintaining a structured computational graph, ensuring gradients flow specifically back to each region's parameters. This can reduces peak memory usage, enabling training with more regions on limited hardware. Standard implementations store attention weights and activations across all heads within a block, and scaling to R^2 heads quickly overwhelms memory. RABs instead keep memory growth roughly $O(RT^2)$ per block.

Segregation of information in attention block supports interpretability. As proposed below, NATs compute attention scores for each target-source regions - making module-level comparisons interpretable and connecting observed dynamics to anatomical origins. This enables for the interpretation of all pair-wise area interactions deeper into the architecture of our model.

Segregation of information at whole-transformer level. While NATs support region-specific attention score computation - when coupled of similarly structured FFNs (see FFN discussion below) - they can also preserve interpretability at the whole-transformer level (see, e.g. Fig 3(a)). This potentially makes it possible to create multi-layer NATs while maintaining separability between region pairs.

2. Approach

We provide a detailed description of NATs that mostly follows the standard transformer (Vaswani et al., 2017b) structure - and explain where our approach differs.

2.1. Neural data representation

We define raw (post-processed) neural data time series from a single source (e.g. brain area, neuron or voxel) as a vector:

$$d \in \mathbb{R}^T.$$

For R sources we can represent the input data as 2-dimensional tensor:

$$D \in \mathbb{R}^{R \times T}.$$

2.2. Embedding and dimensionality expansion

At the embedding stage, each scalar value in the T-long time series is projected into a d_{model}-dimensional embedding space. This *dimensionality expansion* (used in previous work) increases the representational capacity of each neural time point for each area and potentially allows for mixing of information downstream.

Thus, each neural data value (i.e. scalar) becomes a vector:

$$d_t \ \in \ \mathbb{R}^{d_{\text{embed}}}, \quad t = 1, \ldots, T.$$

and for R sources, our embedding is a 3-dimensional tensor:

$$E \ \in \ \mathbb{R}^{R \times T \times d_{\text{embed}}}.$$

That is, each entry $D_{r,t}$ of the raw data tensor is mapped to an embedding vector

$$e_{r,t} \in \mathbb{R}^{d_{\text{embed}}},$$

so that

$$E = \{e_{r,t}\}_{r=1,\ldots,R;\ t=1,\ldots,T}.$$

We note that we do not use positional embedding as part of our approach - as we seek to develop continuous neural time series models (rather than models for event or stimulus-triggered data) and as such - the relative temporal locations of data - do not have specific meaning. This is done without loss of generalization - and a positional embedding step could be added to our structure for specific applications.

2.3. Data flattening

In practice, we preserve the temporal structure while flattening across sources and embedding dimensions. That is, the 3D tensor

$$E \in \mathbb{R}^{T \times R \times d_{\text{embed}}}$$

is reshaped into

$$X \in \mathbb{R}^{T \times (R \cdot d_{\text{embed}})}.$$

and we arrive at standard residual stream size d_{model}:

$$X \in \mathbb{R}^{T \times d_{\text{model}}}, \quad d_{\text{model}} = R \cdot d_{\text{embed}}.$$

Here, the dimension T corresponds to the sequence length. In analogy to large language models (LLMs), it serves as the *context window* (or token window) during training and inference, i.e. the maximum number of time steps the model can jointly attend to.

2.4. Attention block

Commonly used single-head transformer architectures are based on an attention block which projects X onto *Query*, *Key*, and *Value* matrices:

$$Q = XW_Q, \quad K = XW_K, \quad V = XW_V,$$

with learned projection matrices $W_Q, W_K, W_V \in \mathbb{R}^{d_{model} \times d_k}$. And *self-attention* would be defined as:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right) V,$$

where the softmax acts row-wise to normalize over sequence positions.

2.5. Multi-Head Attention

Multi-head attention (MHA) is a commonly used architecture for further expanding the capacity of the attention block. Given the sequence input

$$X \in \mathbb{R}^{T \times d_{\text{model}}}, \tag{1}$$

multi-head attention splits the feature dimension into h heads of size $d_{\text{head}} = d_{\text{model}}/h$.

For each head $i = 1, \ldots, h$ we compute

$$\begin{aligned} Q_i &= XW_i^Q, \quad W_i^Q \in \mathbb{R}^{d_{\text{model}} \times d_{\text{head}}}, \\ K_i &= XW_i^K, \quad W_i^K \in \mathbb{R}^{d_{\text{model}} \times d_{\text{head}}}, \\ V_i &= XW_i^V, \quad W_i^V \in \mathbb{R}^{d_{\text{model}} \times d_{\text{head}}}, \end{aligned} \tag{2}$$

so that $Q_i, K_i, V_i \in \mathbb{R}^{T \times d_{\text{head}}}$.

Each head applies scaled dot-product attention across the T time steps:

$$\text{head}_i = \text{softmax}\left(\frac{Q_i K_i^\top}{\sqrt{d_{\text{head}}}}\right) V_i, \quad \text{head}_i \in \mathbb{R}^{T \times d_{\text{head}}}. \tag{3}$$

The head outputs are concatenated and projected back:

$$H = \text{Concat}(\text{head}_1, \ldots, \text{head}_h) \in \mathbb{R}^{T \times (h d_{\text{head}})}, \tag{4}$$

so that:

$$\begin{aligned} \text{MultiHead}(X) = HW^O \in \mathbb{R}^{T \times d_{\text{model}}}, \\ W^O \in \mathbb{R}^{(h d_{\text{head}}) \times d_{\text{model}}}, \end{aligned} \tag{5}$$

where W^O serves as the *fusion projection.*

2.6. NAT attention block

Below we describe the Neuro-Anatomical Attention (NAT) block and Region-Attention-Block (RABs) to replace the standard MHA. We provide a broader discussion on the differences with existing approaches in Appendix A.

2.7. Region-attention block (RAB)

We propose that NATs (i.e. single-layer attention blocks) contain separate region-attention-blocks (RABs) that are

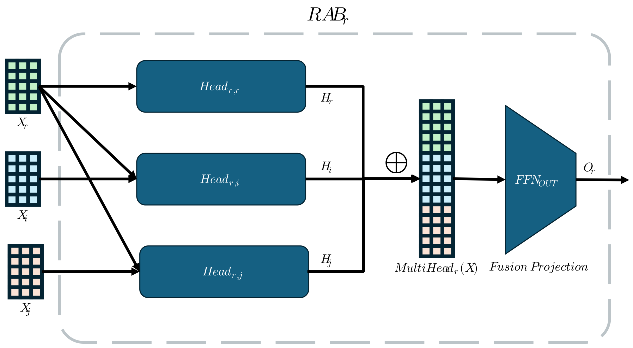

Figure 1. Region-attention-block for single neural area

.

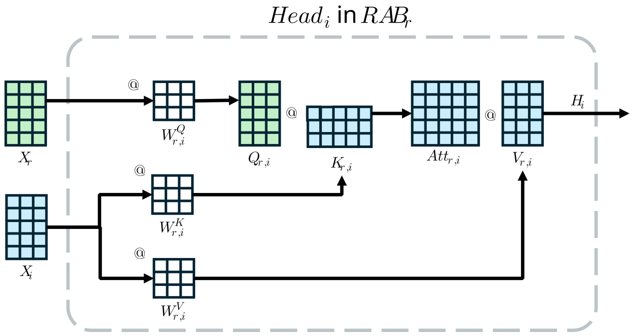

Figure 2. Single Head within RAB (ony 2-neural stream version shown for simplicity)

.

tasked with mixing a single (neural area) time series with all the all the others while preserving interpretability (Fig 1). RABs are constructed for each brain region with the aim of capturing all region–region interactions, while also preserving asymmetry by replicating all *connections* in both directions.

Formally, each RAB corresponds to a single neural region - $r \in [R]$, that has access to both the full embedded input

$$X \in \mathbb{R}^{R \times T \times d},$$

while carrying out a region-specific computation (Fig 1)

$$X_r \in \mathbb{R}^{T \times d}.$$

Thus, within each RAB, we have R heads and each head has unique query matrices as above, i.e.:

$$Q_{r,h} = X_r W^Q_{r,h}, \quad h = 1, \ldots, H,$$

where H denotes the number of attention heads and $W^Q_{r,h} \in \mathbb{R}^{d \times d_h}$ are learned projection matrices (Fig 2). In parallel, for each region $j \in [R]$, we compute region-specific key and value projections:

$$K_{j,h} = X_j W^K_{j,h}, \quad V_{j,h} = X_j W^V_{j,h},$$

with learned matrices $W^K_{j,h}, W^V_{j,h} \in \mathbb{R}^{d \times d_h}$.

Thus, while each query $Q_{r,h}$ originates from the focal region r, the keys and values $(K_{j,h}, V_{j,h})$ are derived from region j. This allows each head in the RAB to align the queries of region r with the representations of a specific region j (including $j = r$).

The attention operation for head h comparing region r to region j is defined as expected:

$$\begin{aligned} &\text{Attention}_{r \to j,h}(Q_{r,h}, K_{j,h}, V_{j,h}) = \\ &\text{softmax}\left(\frac{Q_{r,h} K_{j,h}^\top}{\sqrt{d_h}} \right) \quad V_{j,h}. \end{aligned} \tag{6}$$

where the softmax normalizes across the temporal dimension T.

The outputs of all heads in region r are then concatenated and passed through a learned *fusion projection*:

$$\begin{aligned} \text{RAB}_r(X) = \text{Concat}\Big(&\{\text{Attention}_{r \to j,h}\}_{j=1}^{R}, \\ &h = 1, \ldots, H \Big) \; W^O_r, \end{aligned} \tag{7}$$

with

$$W^O_r \in \mathbb{R}^{(R \cdot H \cdot d_h) \times d}. \tag{8}$$

2.8. NAT feed forward NN architectures

We complete the NAT architecture with two architectural options: a region preserving FFN module (Fig 3a); and (b) a more common FFN module that mixes all signals (Fig 3b).

We provide results in the next section on the effect of the two FFN architectures.

3. Results

We evaluated several architectures derived from the proposed design. As a baseline, we trained a standard 1-layer Transformer with multi-head attention (MHA) and a standard (shared) feed-forward network. We also trained NAT-based models composed of region-attention-blocks (RABs), with two variants of the feed-forward module: (a) region-preserving FFNs, in which each RAB is paired with its own local feed-forward network; and (b) a global FFN, where outputs of all RABs are mixed by a single shared feed-forward network (Fig. 3).

Figure 4 compares validation loss trajectories for baseline MHA models and our proposed NAT architectures. Among

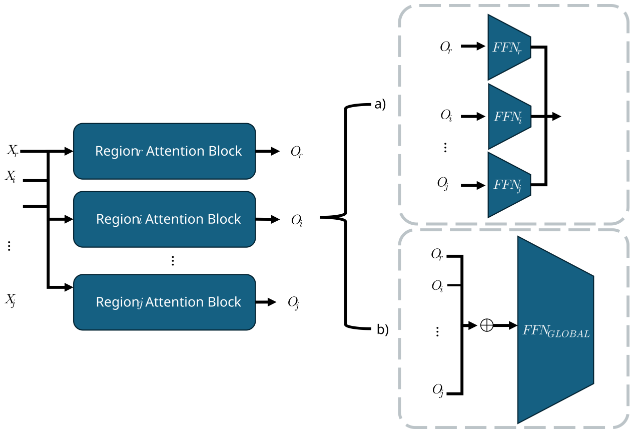
X_r
X_i
X_j
Region r Attention Block
Region i Attention Block
Region j Attention Block
O_r
O_i
O_j
a)
b)
FFN_r
FFN_i
FFN_j
FFN_{GLOBAL}

Figure 3. **NAT feed-forward-nn**.TEXT

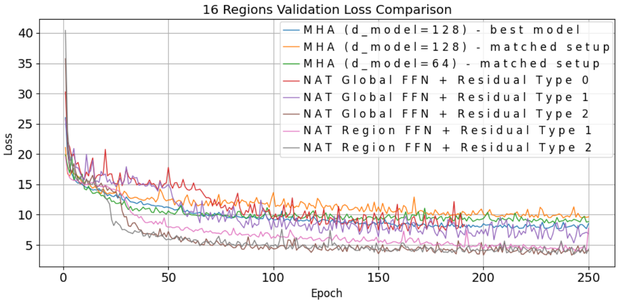

Figure 4. **Validation loss comparison across baseline and NAT models.** Three multi-head attention (MHA) baselines are shown: (i) the best empirically tuned configuration, similar to prior work but with additional modifications, and (ii–iii) two matched setups with $d_{\text{model}} = 128$ and $d_{\text{model}} = 64$, where $d_{\text{hid}} = 4 \cdot d_{\text{model}}$. An attempted run with $d_{\text{model}} = 256$ was excluded due to divergence. NAT variants include both Global and Region FFN modules, each trained with three residual connection strategies: Type 0 (normalized residual added to attention output), Type 1 (unnormalized residual), and Type 2 (unnormalized residual plus an additional FFN residual around the attention output).

the baselines, the empirically tuned MHA model achieves strong performance, while the matched setups with $d_{\text{model}} = 128$ and $d_{\text{model}} = 64$ provide a fairer comparison to NAT.

NAT variants reach lower validation loss across training, supporting that explicit modeling of region-region interactions can improve optimization efficiency and generalization. We provide three types of NATs. Type-0 feed normalized data into the attention layer (similar to standard transformers). Type-1 where we do not normalize the residual stream input to the attention block. Type-2 where we feed non-normalized residual into the attention block output (the attention block score) prior to FFN processing.

We find that Type 0 (normalized residual) remains competitive, while Type 1 (unnormalized residual) improves stability slightly. The strongest performance is obtained with Type 2 - similar to standard transformer residual stream additions into the FFN layer. Our results highlight the flexibility of the NAT design in exploring architectural trade-offs between interpretability and predictive performance.

4. Conclusion

In this work, we introduced the **Neuro-Anatomical Transformer (NAT)**, a transformer-based architecture specifically designed to model large-scale brain dynamics with interpretability, biological realism, and modular structure. Motivated by the growing intersection of neuroscience and AI (NeuroAI), we identified key limitations in current transformer models when applied to neural systems, particularly their lack of alignment with neuroanatomy and limited support for causal or perturbation-driven analysis. NATs address these challenges through modular **Region-Attention Blocks (RABs)**, which explicitly model asymmetric interactions between brain areas and maintain segregated information pathways throughout the network (see Section 2).

Our experiments on widefield calcium imaging data across 16 mouse brain regions (Mitelut et al., 2022) demonstrate that NATs can match or exceed the performance of standard transformers while significantly improving anatomical fidelity and interpretability (see Section 3). These gains come with only modest architectural constraints and enable structured comparisons across region–region attention maps, making it possible to inspect and analyze learned representations with causal precision. We propose that NATs can support *in silico* perturbation experiments and preserve interpretability even in deep, multi-layer transformer stacks when paired with region-aligned feedforward modules (Fig. 3).

Looking ahead, we propose that NATs could serve as a foundation for **safe and interpretable AI systems**, including as a stepping stone toward **whole-brain emulation (WBE)**. As AI systems grow in complexity, there is a pressing need for architectures whose internal dynamics can be understood and manipulated. NATs offer a uniquely promising solution: a white-box transformer framework that bridges neuroscience and alignment research by enabling fine-grained control, hypothesis testing, and failure mode discovery. In future work, we plan to expand NATs to richer multimodal datasets, explore scaling behavior across larger brain architectures, and develop their use as alignment testbeds and cognitive simulators. We believe that biologically grounded architectures like NATs may not only advance our understanding of the brain but also inform the design of transparent, agentic, and safer AI systems.

References

Dong, J., Wu, H., Zhang, H., Zhang, L., Wang, J., and Long, M. Simmtm: A simple pre-training framework for masked time-series modeling. In *Advances in Neural Information Processing Systems*, 2023. URL https://arxiv.org/abs/2302.00861.

Fernando, J. and Guitchounts, G. Transformer dynamics: A neuroscientific approach to interpretability of large language models. *Preprint (arXiv)*, 2025. Models transformer residual streams as evolving dynamical systems for mechanistic interpretability.

Freeman, L., Shamash, P., Arora, V., Barry, C., Branco, T., and Dyer, E. Beyond black boxes: Enhancing interpretability of transformers trained on neural data. *Preprint (arXiv)*, 2025. Transformer + sparse autoencoders applied on calcium imaging for interpretable neural features.

Hou, Z., Liu, X., Cen, Y., Dong, Y., Yang, H., Wang, C., and Tang, J. Graphmae: Self-supervised masked graph autoencoders. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '22)*, pp. 594–604, 2022. doi: 10.1145/3534678.3539321. URL https://arxiv.org/abs/2205.10803.

Kan, X., Dai, W., Cui, H., Zhang, Z., Guo, Y., and Yang, C. Brain network transformer. *Preprint (arXiv)*, 2022. Transformer modeling of brain networks using ROI connectivities for classification.

Kim, P. Y., Kwon, J., Joo, S., Bae, S., Lee, D., Jung, Y., Yoo, S., Cha, J., and Moon, T. Swift: Swin 4d fmri transformer. In *Preprint (arXiv)*, 2023. 4D Swin Transformer for fMRI analysis with explainable attention maps.

Kipf, T., Fetaya, E., Wang, K.-C., Welling, M., and Zemel, R. Neural relational inference for interacting systems. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 2688–2697, Stockholm, Sweden, 10–15 Jul 2018. PMLR. URL https://proceedings.mlr.press/v80/kipf18a.html.

Li, Y., Yu, R., Shahabi, C., and Liu, Y. Diffusion convolutional recurrent neural network: Data-driven traffic forecasting. In *International Conference on Learning Representations (ICLR)*, 2018. URL https://openreview.net/forum?id=SJiHXGWAZ.

Li, Z., Rao, Z., Pan, L., Wang, P., and Xu, Z. Ti-mae: Self-supervised masked time series autoencoders. *arXiv preprint arXiv:2301.08871*, 2023. URL https://arxiv.org/abs/2301.08871.

Mitelut, C., Zhang, Y., Sekino, Y., Boyd, J. D., Bollanos, F., Swindale, N. V., Silasi, G., Saxena, S., and Murphy, T. H. Mesoscale cortex-wide neural dynamics predict self-initiated actions in mice several seconds prior to movement. *eLife*, 11:e76506, 2022. doi: 10.7554/eLife.76506.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 5998–6008, 2017a.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u., and Polosukhin, I. Attention is all you need. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017b. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

Velickovic, P., Cucurull, G., Casanova, A., Romero, A., Liò, P., and Bengio, Y. Graph attention networks. In *International Conference on Learning Representations (ICLR)*, 2018. URL https://arxiv.org/abs/1710.10903.

Wu, Z., Pan, S., Long, G., Jiang, J., and Zhang, C. Graph wavenet for deep spatial-temporal graph modeling. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence (IJCAI-19)*, pp. 1907–1913, 2019. doi: 10.24963/ijcai.2019/264. URL https://www.ijcai.org/proceedings/2019/264.

Ye, J. and Pandarinath, C. Neural data transformer 2 (ndt2): Multi-context pretraining for neural spiking activity. *Preprint*, 2023. Demonstrates spatiotemporal transformer pretraining across sessions, subjects, and tasks.

A. Related Work

B. Related Work

B.1. Diffusion-based graph priors

Graph-based time series models frequently encode inter-area structure with a fixed adjacency matrix $A \in \mathbb{R}^{R \times R}$. The Diffusion Convolutional Recurrent Neural Network (DCRNN) models spatial coupling as diffusion over directed random walks and integrates this with recurrent temporal dynamics (Li et al., 2018). Its core operator applies bidirectional K-step diffusion filters to region features $X_t \in \mathbb{R}^{R \times d}$, thereby propagating information along graph edges before updating hidden states with gated recurrent units.

B.2. Adaptive adjacency and spatio-temporal convolutions

Graph WaveNet extends this idea by learning an adaptive adjacency matrix in addition to using the fixed structural graph (Wu et al., 2019). Node embeddings are optimized jointly with the model to produce a dynamic adjacency $\tilde{A}_{\text{adp}}$, which is combined with diffusion powers and dilated one-dimensional convolutions. This allows the model to capture both long-range temporal dependencies and flexible, data-driven inter-area interactions.

B.3. Graph-masked attention

Another line of work applies attention mechanisms constrained by graph structure. Graph Attention Networks (GAT) restrict attention to a node's neighbors $\mathcal{N}_A(r)$ and learn weights α_{ij} that modulate information aggregation (Velickovic et al., 2018). Multi-head attention layers then allow parallel subspace interactions, while the graph mask enforces locality and interpretability. This approach has been widely adopted in spatio-temporal brain modeling, where neighborhood structure is biologically motivated.

B.4. Latent graph inference

Some approaches infer the interaction graph itself as a latent variable. Neural Relational Inference (NRI) introduces a variational autoencoder that simultaneously learns a latent edge distribution and system dynamics (Kipf et al., 2018). A GNN encoder infers discrete interaction types between nodes, while a GNN decoder predicts future trajectories given sampled edges. This formulation captures directed, sparse, and potentially dynamic interaction patterns without requiring a fixed prior adjacency.

B.5. Masked pretraining for time series and graphs

Recent work has also explored masked reconstruction objectives as self-supervised pretraining. For time series, models such as Ti-MAE and SimMTM mask temporal segments and reconstruct the missing values to encourage robust representations (Li et al., 2023; Dong et al., 2023). In the graph domain, GraphMAE reconstructs masked node features using a scaled cosine error, encouraging embeddings to capture structural context (Hou et al., 2022). These methods inject inductive biases by training models to recover missing information under structured masking.

B.6. Comparison with NAT

Compared to these prior approaches, Neuro-Anatomical Transformers (NATs) aim to learn region–region relationships directly from task loss, without relying on auxiliary objectives or fixed priors. Like GAT and NRI, they are capable of modeling directed and asymmetric interactions, but they do so through region-conditioned attention rather than graph message passing or latent-variable inference. NATs can incorporate external priors in a manner similar to graph-based methods (e.g., masking logits with Π), yet their core design is agnostic to explicit graph structure. This makes them a flexible and interpretable alternative: they retain the expressive capacity of multi-head attention while maintaining a neuro-anatomical decomposition of interactions.