Building Interpretable, Trust-worthy Systems for Neural Signal Decoding

Hua Xu[1]

[1]The Hong Kong University of Science and Technology (Guangzhou)
hxu401@connect.hkust-gz.edu.cn

Abstract

While deep learning excels at decoding neural signals, the opacity of state-of-the-art models limits their scientific utility and clinical trustworthiness. We propose a research that bridges this gap by integrating high-performance architectures—specifically Transformers and Graph Neural Networks—with mechanistic interpretability and neuro-symbolic reasoning. This proposal aims to uncover verifiable mappings between artificial computational circuits and biological dynamics without compromising decoding accuracy. Validated through rigorous benchmarking and wet-lab experiments, this work establishes a foundation for transparent brain-computer interfaces and accelerates fundamental neuroscience research.

1 Introduction

My proposed research addresses a central challenge at the intersection of neuroscience and artificial intelligence: building interpretable and trustworthy AI systems for decoding neural signals. Neural activity, captured through high-dimensional modalities like Magnetoencephalography (MEG), Electroencephalography (EEG), and functional MRI (fMRI), holds the key to understanding cognition and treating neurological disorders. Accurately interpreting these complex patterns is essential for unlocking the mechanisms behind mental health conditions and enabling transformative technologies. However, a critical bottleneck remains. While deep learning models have achieved impressive decoding performance, their "black box" nature limits their utility. Neuroscientists require more than just accurate predictions; they need to uncover the biological mechanisms driving those predictions—the "why" behind the neural phenomena. Furthermore, clinical applications demand transparency. Technologies like BCIs require systems whose decision-making processes are fully understood to ensure safety and establish user trust. To address this, I propose developing interpretable models specifically designed for the hierarchical and multi-modal nature of neural data. This project will extend mechanistic interpretability techniques beyond their current focus on language models, pioneering new methods to make neural decoders transparent. By creating a system that synergizes predictive power with explanatory depth, this work aims to bridge the gap between AI capabilities and scientific needs, establishing a foundation for trustworthy neuro-technologies that can safely transition from the laboratory to real-world application.

2 Background

2.1 Related Work

Research in neural signal decoding has progressed along two parallel tracks: one prioritizing predictive performance and the other pursuing interpretability. The performance-focused track employs deep learning architectures, including large-scale foundation (Qiu et al. 2025) and generative models (Luo et al. 2024), to map neural recordings (fMRI, EEG) to behavioral outcomes or stimuli (Gokce and Schrimpf (2024); Dai et al. (2025)). While these models achieve SOTA accuracy, their increasing complexity renders them opaque "black boxes" that fail to explain their underlying computational processes. Concurrently, the field of mechanistic interpretability has matured for Large Language Models (LLMs), yielding powerful techniques to identify functional computational units (Lindsey et al. 2025) or computational phenomenon (Cunningham et al. (2023); nostalgebraist (2020)) within artificial neural networks. However, these advanced methods are rarely applied in computational neuroscience. Instead, existing interpretability efforts in this domain typically rely on either neuro-symbolic methods (Castro et al. 2025), which can become unwieldy when generating massive amounts of explicit formulae, or overly simplistic models (Li, Benna, and Mattar 2024) that lack the capacity to generalize. This project is situated at the confluence of these tracks, aiming to bridge this gap by applying advanced mechanistic interpretability techniques to high-performance neural decoding models. The goal is to create a system that is both powerful and transparent, avoiding the conventional trade-off between predictive accuracy and scientific insight.

2.2 Prior Work by the Applicant

My prior research provides a strong foundation for this proposed work in two key areas. **Multivariate Irregular Time Series Forecasting**. Previously, I co-developed a novel method for irregularly sampled time series (Hu et al. 2025), a common format for neural data. By adapting a pretrained visual masked autoencoder to convert sparse data into image-

like patches, we effectively captured cross-channel dependencies and achieved strong few-shot performance. This experience directly prepares me for the challenges of modeling complex neural signals.

Rule-based Reasoning on LLMs. I have worked on a system that enhances an LLM's ability to generate interpretable scientific rules from data (Yang et al. 2025). By integrating probabilistic methods with LLM-based rule generation, we improved both performance and the novelty of the resulting hypotheses. This work is highly relevant to my goal of extracting abstract, verifiable rules from neural decoding models to enhance their interpretability.

3 Approach

This research requires efforts in two domains, building powerful neural signal decoding models, and interpreting built AI systems.

3.1 Developing a High-Performance Neural Decoder

A prerequisite for a meaningful interpretability analysis is a model that demonstrates high performance on core decoding tasks. I will construct a powerful base model by drawing from and potentially hybridizing advanced architectures.

To capture the complex temporal dynamics and long-range dependencies inherent in neural time-series data, transformer-based backbones will be used to ensure performance. Spatial relationships between brain regions, GNN may be integrated into the architecture. For functional decoding tasks such as reconstructing sensory stimuli from neural data, we will employ powerful generative models with a particular focus on diffusion models, which have shown exceptional performance in high-fidelity generation.

3.2 Integrating a Suite of Interpretability Methods

To "open the black box" of the high-performance decoder, I will apply a toolkit of interpretability techniques drawn from complementary areas of AI research. **Mechanistic Interpretability**. I will use SOTA techniques such as logit lens, sparse autoencoders and circuit analysis. These methods will be used to decompose the model's internal representations and identify semantically meaningful computational circuits—for example, linking specific artificial neuron activations to distinct features of a stimulus. **Neuro-Symbolic Reasoning**. To make the model's operations human-readable, I will employ neuro-symbolic methods to translate the identified circuits or decision pathways into explicit mathematical formulas or logical rules. This creates a clear, verifiable mapping from input to output.

Advanced Mathematical Analysis. Inspired by theoretical neuroscience, which uses tools like manifold and group theory to describe dynamics of biological neural networks, I will apply these mathematical lenses to analyze the geometry of the representations learned by the artificial decoder. This can reveal principles about how models organize and process information. By synthesizing these two components, this research will create a framework where the decoding model can provide accurate forecasting abilities while the interpretable modules can help to make scientific discoveries, making the framework powerful and transparent.

4 Evaluation

The system's success will be evaluated through a three-tiered approach assessing its performance, the utility of its interpretations, and its real-world scientific validity.

Quantitative Benchmarking. The model's predictive accuracy on tasks like signal forecasting will be benchmarked against state-of-the-arts non-interpretable models using standard metrics (e.g., MSE). Success is defined as achieving comparable performance, proving that interpretability does not significantly compromise accuracy.

Qualitative User Studies. In collaboration with neuroscientists, the system's interpretations will be evaluated for clarity, biological plausibility, and their potential to generate novel hypotheses. The system will be considered successful if domain experts consistently rate its outputs as useful for inspiring new research.

Experimental Validation. The ultimate test will be to experimentally validate a novel, model-generated hypothesis in a wet-lab setting (e.g., via optogenetic stimulation), which would provide definitive evidence that the system can serve as a genuine tool for scientific discovery.

5 Discussion

This research is expected to discover a verifiable mapping between a model's internal computations and the brain's biological processes, providing a mechanistic understanding of neural data processing. Such a breakthrough would equip neuroscientists with a powerful tool for hypothesis generation while pioneering new AI interpretability techniques for complex data. Ultimately, by creating transparent and reliable models, this work will build the trust needed to develop safer clinical applications, such as advanced diagnostics and next-generation brain-computer interfaces (BCIs).

6 Conclusion

This proposal addresses the critical "black box" nature of neural decoding models, which hinders both scientific discovery and safe clinical application. I will develop a system that synergizes a high-performance deep learning model with a suite of mathematical, mechanistic and neuro-symbolic interpretability tools. Its success will be validated through a rigorous, multi-tiered evaluation, from computational benchmarks to the experimental confirmation of model-generated hypotheses. This work will pioneer methods for transparent, science-aligned AI, building the trust necessary to translate technologies like brain-computer interfaces from the lab into real-world applications that improve human well-being.

References

Castro, P. S.; Tomasev, N.; Anand, A.; Sharma, N.; Mohanta, R.; Dev, A.; Perlin, K.; Jain, S.; Levin, K.; Elteto, N.; Dabney, W.; Novikov, A.; Turner, G. C.; Eckstein, M. K.; Daw,

N. D.; Miller, K. J.; and Stachenfeld, K. 2025. Discovering Symbolic Cognitive Models from Human and Animal Behavior. In *Forty-second International Conference on Machine Learning*.

Cunningham, H.; Ewart, A.; Riggs, L.; Huben, R.; and Sharkey, L. 2023. Sparse Autoencoders Find Highly Interpretable Features in Language Models. arXiv:2309.08600.

Dai, Y.; Yao, Z.; Song, C.; Zheng, Q.; Mai, W.; Peng, K.; Lu, S.; Ouyang, W.; Yang, J.; and Wu, J. 2025. MindAligner: Explicit Brain Functional Alignment for Cross-Subject Visual Decoding from Limited fMRI Data. arXiv:2502.05034.

Gokce, A.; and Schrimpf, M. 2024. Scaling Laws for Task-Optimized Models of the Primate Visual Ventral Stream. arXiv:2411.05712.

Hu, Z.; Wu, J.; Xu, H.; Liao, M.; Feng, N.; Gao, B.; Lai, S.; and Yue, Y. 2025. IMTS is Worth Time $\times$ Channel Patches: Visual Masked Autoencoders for Irregular Multivariate Time Series Prediction. arXiv:2505.22815.

Li, J.-A.; Benna, M. K.; and Mattar, M. G. 2024. Discovering cognitive strategies with tiny recurrent neural networks. *Nature*, 644: 993 – 1001.

Lindsey, J.; Gurnee, W.; Ameisen, E.; Chen, B.; Pearce, A.; Turner, N. L.; Citro, C.; Abrahams, D.; Carter, S.; Hosmer, B.; Marcus, J.; Sklar, M.; Templeton, A.; Bricken, T.; McDougall, C.; Cunningham, H.; Henighan, T.; Jermyn, A.; Jones, A.; Persic, A.; Qi, Z.; Thompson, T. B.; Zimmerman, S.; Rivoire, K.; Conerly, T.; Olah, C.; and Batson, J. 2025. On the Biology of a Large Language Model. *Transformer Circuits Thread*.

Luo, A. F.; Yeung, J.; Zawar, R.; Dewan, S.; Henderson, M. M.; Wehbe, L.; and Tarr, M. J. 2024. Brain Mapping with Dense Features: Grounding Cortical Semantic Selectivity in Natural Images With Vision Transformers. *arXiv preprint arXiv:2410.05266*.

nostalgebraist. 2020. Interpreting GPT: the logit lens. https://www.lesswrong.com/posts/AcKRB8wDpdaN6v6ru/interpreting-gpt-the-logit-lens. Accessed: 2025-02-22.

Qiu, W.; Huang, Z.; Hu, H.; Feng, A.; Yan, Y.; and Ying, R. 2025. MindLLM: A Subject-Agnostic and Versatile Model for fMRI-to-Text Decoding. arXiv:2502.15786.

Yang, Y.; Xu, H.; Hu, Z.; and Yue, Y. 2025. RLIE: Rule Generation with Logistic Regression, Iterative Refinement, and Evaluation for Large Language Models. arXiv:2510.19698.