

Principled Neuroscientific Discovery with Machine Learning

By

Alicia Zeng

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Biophysics

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Jack Gallant, Chair

Professor Bruno Olshausen

Professor Frederic Theunissen

Professor Richard Ivry

Fall 2025

Principled Neuroscientific Discovery with Machine Learning

Copyright 2025
by
Alicia Zeng

Abstract

Principled Neuroscientific Discovery with Machine Learning

by

Alicia Zeng

Doctor of Philosophy in Biophysics

University of California, Berkeley

Professor Jack Gallant, Chair

Understanding brain function requires modeling a high-dimensional, nonlinear dynamical system continuously driven by sensory input, shaped by internal state and context, and expressed through behavior. Machine learning can be used to extract predictive structure from large-scale neural recordings, but its scientific value hinges on whether fitted parameters admit interpretation as testable hypotheses about neural computation. This thesis develops and applies interpretable predictive models—primarily linearized encoding models that map stimulus or behavioral features onto neural responses—to investigate (i) how human cortex represents abstract concepts during naturalistic language comprehension and (ii) how deep brain stimulation (DBS) modulates movement-related neural dynamics in Parkinson’s disease.

The first study addresses a central limitation of encoding models built on dense word embeddings or neural network activations: superposition. When latent semantic factors outnumber embedding dimensions, distinct concepts become entangled along shared directions, rendering voxelwise regression weights uninterpretable and formally non-identifiable. I introduce the Sparse Concept Encoding Model, which applies sparse dictionary learning to project dense embeddings into an overcomplete basis of interpretable concept atoms. Because each atom isolates a coherent semantic direction, voxel tuning can be read directly from regression coefficients. Applied to whole-brain fMRI acquired during naturalistic story listening, this approach matches the prediction accuracy of conventional dense models while substantially improving interpretability. The resulting cortical maps reveal convergent selectivity for time, space, and number in the intraparietal sulcus and the

inferior frontal gyrus—consistent with the shared-magnitude hypothesis—and uncover systematic similarity structure among distributed semantic representations.

The second study moves to a clinical domain where characterizing neural dynamics carries immediate therapeutic implications. Parkinson’s disease is increasingly understood as a disorder of network dynamics wherein motor symptoms arise from aberrant oscillatory activity across basal ganglia and sensorimotor cortex. Yet, DBS remains an open-loop therapy with parameters tuned heuristically. Using a home-deployed multi-modal data collection platform integrating bidirectional neurostimulators, bilateral wrist accelerometry, and video-based pose estimation, we recorded synchronized neural and behavioral data from two participants performing standardized motor tasks across multiple stimulation amplitudes. By building encoding models that predict neural spectral power data from behavioral features, we demonstrate how temporally rich predictors such as full acceleration spectrograms and task labels outperform coarse movement summaries. Thus, this study highlights the importance of fine-grained behavioral structure for identifying movement-related variance in sensed neural signals. In the participant with stronger baseline model performance, higher stimulation amplitudes enhanced prediction performance across sensorimotor cortex and subthalamic nucleus. We show that these gains arise from amplification of a movement-related spectral component characterized by beta suppression alongside alpha/delta and gamma enhancement, including a prominent peak at the stimulation-entrained subharmonic. This interpretation is consistent with “information lesion” accounts of DBS, wherein pathological synchronization impedes movement-related signaling and high-frequency stimulation restores information flow through basal ganglia–cortical circuits.

Together, these studies illustrate how high-volume data collected under naturalistic or semi-naturalistic conditions can be analyzed with interpretable predictive models to transform accurate prediction into mechanistic neuroscientific hypotheses.

Dedicated to my grandma, 赵友庄.

Contents

Acknowledgments	viii
1 Introduction: Principled neuroscientific discovery with machine learning	1
1.1 Interpretable Brain Encoding Model with Sparse Concept Atoms	2
1.2 Deep Brain Stimulation Amplifies Movement-Related Neural Modulation in Parkinson's Disease	3
1.3 Summary and Motivation	4
2 Disentangling Superpositions: Interpretable Brain Encoding Model with Sparse Concept Atoms	6
2.1 Abstract	6
2.2 Introduction	6
2.3 Background	9
2.3.1 Sparse Coding	9
2.3.2 Sparse Coding For Interpreting Word Embeddings and Deep Neural Networks	9
2.4 Theory	10
2.4.1 Limitations of Dense Encoding Models	10
2.4.2 Theorem (Non-identifiability in Dense Embedding Spaces)	12
2.4.3 The Sparse Concept Encoding Model	13
2.4.4 Bayesian Interpretation of Sparse and Dense Encoding Models	13
2.5 Methods	15
2.5.1 Experimental Procedure	15

2.5.2	Dataset Structure and Test/Train Split	15
2.5.3	fMRI Preprocessing	16
2.5.4	Feature Construction	16
2.5.5	Feature Normalization	16
2.5.6	Model Estimation	17
2.5.7	Model Evaluation	17
2.5.8	Postprocessing of Semantic Weights	17
2.5.9	Top 50 Sparse Concept Atoms and Associated Words	18
2.5.10	Semantic Category Selection for Time, Number, and Space	18
2.5.11	Software Implementation	19
2.6	Results	19
2.6.1	Voxelwise Modeling and Feature Construction	19
2.6.2	Improving Embedding Geometry with Vector Norm Reparameteri- zation	21
2.6.3	Comparing Predictive Performance of Dense and Sparse Models	21
2.6.4	Concept-Level Cortical Maps from the Sparse Concept Encoding Model	21
2.6.5	Case Study: Disentangling Time, Space, and Number Representations	23
2.6.6	Case Study: Comparing Cortical Representations of Different Concepts	26
2.7	Discussions	26
2.8	Supplementary Materials	28
2.8.1	Heuristic Concept-Level Cortical Maps from the Dense Encoding Model	30
2.8.2	Statistical Significance Maps from Permutation Testing	31
2.8.3	Evaluation of Embedding Preprocessing Methods	33
3	Deep Brain Stimulation Amplifies Movement-Related Neural Modulation in Parkinson's Disease: Evidence from Multimodal Home Recordings	36
3.1	Abstract	36
3.2	Introduction	37
3.3	Background	38
3.3.1	Parkinson's Disease and Deep Brain Stimulation	38
3.3.2	Hypothetical Mechanisms of DBS	39
3.3.3	Limitations of Current DBS and Directions for Improvement	41
3.3.4	Neural Biomarkers for Parkinson's Disease	42
3.3.5	Bidirectional Implantable Neurostimulators	42
3.4	Methods	43
3.4.1	Experimental Design	43
3.4.2	Data Collection Platform	44
3.4.3	Stimulation Amplitude Titration	46

3.4.4	RCS Time Series Preprocessing	47
3.4.5	Neural Spectral Feature Extraction	47
3.4.6	Watch Accelerometer Spectral Feature Extraction	48
3.4.7	Task Label Feature Extraction	48
3.4.8	Pose Feature Extraction	48
3.4.9	Encoding Model Construction	49
3.4.10	Feature Preprocessing	50
3.4.11	Model Estimation	50
3.4.12	Model Evaluation	51
3.4.13	Statistical Analysis	51
3.5	Results	53
3.5.1	Training Multimodal Encoding Models	53
3.5.2	Evaluating Encoding Model Prediction Performances	57
3.5.3	Stimulation Amplitude Modulates Encoding Model Performances	61
3.5.4	Stimulation Amplifies a Movement-Locked Cortical Component	65
3.6	Discussion	68
3.6.1	Limitations	69
3.6.2	Future Directions	71
3.7	Supplementary Materials	71
3.7.1	Patient Instructions	71
4	Conclusion	74
4.1	Summary of Contributions	74
4.1.1	Chapter 2: Interpretable Brain Encoding Model with Sparse Concept Atoms	74
4.1.2	Chapter 3: Deep Brain Stimulation Amplifies Movement-Related Neural Modulation in Parkinson's Disease	75
4.2	Limitations and Lessons	75
4.3	Looking Forward	76
	References	78

List of Figures

2.1	The Sparse Concept Encoding Model resolves superposition in dense feature spaces.	8
2.2	Sparse dictionary learning decomposes each word into a sparse, non-negative linear combination of concept atoms.	10
2.3	The Sparse Concept Encoding Model produces interpretable voxel weights by replacing dense features with learned concept atoms.	11
2.4	The Sparse Concept Encoding Model preserves prediction accuracy while improving interpretability.	22
2.5	The Sparse Concept Encoding Model yields consistent, interpretable cortical maps for individual semantic concepts.	24
2.6	Sparse Concept Encoding Model reveals overlapping cortical representations of time, space, and number.	25
2.7	Sparse Concept Encoding Model reveals consistent structure in concept-level cortical representations across subjects.	27
2.8	Heuristic concept-level cortical maps from the Dense Encoding Model for concept atoms 456 and 115	30
2.9	Heuristic composite cortical maps from the Dense Encoding Model for three concept atoms representing number, time, and space.	32
2.10	Binary statistical significance maps of prediction accuracy for the Sparse Concept Encoding Model	33
2.11	Effect of Vector Norm Reparameterization (VNR) on embedding norms	34

3.1	Multimodal recordings across stimulation conditions, Patient 01	54
3.2	Multimodal recordings across stimulation conditions, Patient 02	56
3.3	Behavioral features predict neural spectral activity at sensing contacts	59
3.4	Stimulation amplitude modulates encoding model performance	63
3.5	Stimulation amplitude modulates encoding model performance in canonical oscillation bands across regions and patients	65
3.6	Stimulation amplifies a movement-related cortical component	67

List of Tables

2.1	Fifty concept atoms with the highest average cortical activation, listed with their top associated words.	20
2.2	Concept atoms with strongest activation for semantic categories of time, number, and space.	29
2.3	Comparison of preprocessing strategies for word embeddings prior to sparse coding	35

Acknowledgments

I am deeply grateful to the many awe-inspiring people who have illuminated the path leading to this thesis.

As an undergraduate, Professor Jeffery Tlumak supervised my unconventional philosophy honors thesis on AlphaGo, written as the program defeated the human world champion in my junior year. Developing that thesis, *Go Metaphors: Meaning, Mind and Machines*, convinced me that rich, meaningful structure could emerge from simple computational elements—an intuition that set me on the trajectory continuing here. In Advanced Quantum Mechanics, Professor Alfredo Gurrola took me and my only classmate to lunch, astonished us with the elegance of physics, and insisted that we were capable of doing research. Douglas Hofstadter welcomed me into his home at a moment when I was profoundly uncertain about my path, and his way of thinking in *Gödel, Escher, Bach* has always accompanied me. Later, Professor Vijay Balasubramanian took me on as a postbaccalaureate researcher in his Physics of Living Matter lab at the University of Pennsylvania; his relentless intellectual curiosity and versatility were both humbling and contagious, and he encouraged me to pursue a PhD at a time when I still doubted that I belonged in one.

At Berkeley, I was fortunate to join Professor Jack Gallant's lab at what was, for me, a demanding moment in life, having just become a new parent. Jack created and maintains a lab culture that strives for the highest standards of scientific rigor and embodies a willingness to question, criticize, and cross-examine one's own thinking without mercy. I am especially grateful for the two four-hour qualifying-exam practice sessions and the many lab meetings in which my presentations were torn apart so that they could be rebuilt stronger. The numerous, often heated, arguments with Jack reshaped how I think about pragmatism, about empiricism versus rationalism, and about the social and political structure of scientific research. Previous and current lab members—including Alex Huth, James Gao, Tom Dupré la Tour, Anwar Nunez-Elizalde, and Matteo Visconti di Oleggio Castello—spent years building the methodological foundations that made the fMRI-based chapter of this thesis possible, and set a very high bar for taking risks in pursuit of genuinely new ideas. Beyond Jack's distinctive blend of irreverence and insight, I was drawn to the lab for its commitment to naturalistic experiments, high-volume data collection, predictive modeling, and a system-identification approach to understanding the brain as a high-dimensional nonlinear system. The Gallant Lab has been developing these

ideas for more than two decades, and recent shifts in scientific practice toward machine learning on large datasets only underscore the prescience of that vision.

A second important intellectual home at Berkeley was the Redwood Center for Theoretical Neuroscience, and especially Professor Bruno Olshausen. Bruno’s combination of deep ingenuity and striking humility fostered an open, welcoming, and fiercely curious culture. Some of my most illuminating moments came in the High-Dimensional Statistics reading group organized by then-postdoc Anthony Thomas, and in the Bayesian Learning / Sensorimotor AI journal club led by postdoc Hadi Vafaii. The idea for the second chapter of this thesis first took shape during my first-year rotation in Bruno’s lab, listening to Yubei Chen describe sparse manifold transforms and sparse coding applied to large language models. Yubei’s work, which later helped inspire industry efforts on sparse autoencoders for mechanistic interpretability, showed me how visionary ideas from theoretical neuroscience could ripple outward and influence the broader trajectory of AI.

I also owe special thanks to each of my thesis and qualification committee members. Professor Richard Ivry co-taught an inspiring Socratic-style seminar on analyzing research papers that fundamentally changed how I interrogate assumptions, methods, and claims in scientific literature. Professor Frédéric Theunissen welcomed me into his lab during the early days of COVID-19, giving me space to explore the use of neural networks in bird vocalization classification. Dr. Simon Little served on my qualification committee and co-led the deep brain stimulation project; his clinical insight and curiosity helped bridge the gap between abstract models and patients’ lived experience. I am grateful to each of them for their warmth, their candor, and the many invigorating conversations from which I learned more than I realized in the moment.

During my time at Berkeley, I was also privileged to follow my curiosity into classrooms across campus. Courses such as *Designing, Visualizing and Understanding Deep Neural Networks* (Anant Sahai), *Convex Optimization (in the Age of LLMs)* (Benjamin Recht), *Neural Computation* (Bruno Olshausen), *Principles of Magnetic Resonance Imaging* (Michael Lustig), and *Computational Imaging* (Laura Waller) opened doors to new ways of thinking and left a lasting imprint on how I approach scientific problems. Numerous seminars at the Simons Institute were equally formative, especially the “AI $\equiv$ Science” series organized by Aditi Krishnapriyan and Jennifer Listgarten, which offered a vivid glimpse of how machine learning and the natural sciences might evolve together.

Looking ahead, I can only hope to carry forward the torches lit by these remarkable minds and to keep that light burning for those who come next.

CHAPTER 1

Introduction: Principled neuroscientific discovery with machine learning

Artificial intelligence has undergone an extraordinary transformation over the span of my PhD. Around 2016–2017, AlphaGo defeated human masters at Go, a 2,500-year-old game whose combinatorial space is so vast that exhaustive search is impossible in practice. Around the same time, the Gallant Lab began using word embeddings in their encoding model pipeline to predict brain activity. These embeddings represent words as points in a geometric space, where distances and directions reflect relationships in meaning. In doing so, they evoke Ludwig Wittgenstein’s idea that ‘the meaning of a word is its use in the language,’ turning patterns of co-occurrence in natural language into geometry in a learned vector space.

These developments were what drew me into this field. Yet when I began my PhD in 2020, even against this backdrop, almost no one I spoke with in adjacent research areas believed that language models would approach human-level language abilities in the near term. Tasks that were widely regarded as out of reach, from writing usable code or solving Olympiad-level mathematics problems to engaging in extended technical dialogues, became routine capabilities of large models within just a few years. As I complete this work, it feels as though I started my PhD in one technological epoch and finished it in another.

For me and many others, the most compelling promise of machine learning has always been its potential to advance the sciences. Treating neurological disorders, curing cancers, and discovering new materials are more than technical milestones; they are breakthroughs that could profoundly reshape how we live and reduce suffering. They are also precisely the domains where unaided human intuition is weakest: the systems involved are noisy, high-dimensional, nonlinear, and often far from equilibrium.

In response to this promise, many groups, including large industrial labs and research institutes, are now pursuing the development of “AI scientists”: language model–centric systems that read papers, synthesize literature, and generate hypotheses. Such tools may be valuable, particularly for organizing and sharing knowledge. Yet in most scientific domains they cannot address the principal bottleneck. For complex systems in molecular biology, neuroscience, and materials science, the true constraint is not the number of ideas that can be written down but the availability of rich, high-quality experimental data against which to evaluate these ideas. These systems inhabit enormous state spaces; the set of hypotheses humans have articulated and tested is a measure-zero subset of what is possible. Genuine progress requires dense, carefully designed measurements in the environments where these systems actually operate, rather than just more hypotheses on a page.

Machine learning models are, at their core, flexible high-dimensional nonlinear functions. With the right data, they can detect patterns, uncover latent structure, and suggest hypotheses at scales unachievable with human cognition alone. But without such data, and without an inductive structure to keep the models interpretable, they risk becoming just another opaque layer between scientists and the underlying phenomena. The central question for “AI for science” is therefore not simply how powerful our models are, but how to couple these models to experiments so that they enable understanding rather than mere prediction. This thesis presents two projects that move deliberately toward a more data-grounded and principled vision of machine learning for science.

1.1 Interpretable Brain Encoding Model with Sparse Concept Atoms

In Chapter 2, adapted from (Zeng and Gallant, 2025), I ask how the brain represents abstract ideas such as time, space, and number, using fMRI encoding models as the main tool. In this framework, a naturalistic stimulus is transformed into a feature matrix, and a regularized linear model is fit to predict the corresponding pattern of brain responses. Each voxel is thus associated with a weight vector, a direction in feature space that, in principle, reveals which aspects of the stimulus that voxel is tuned to. The central difficulty lies in designing this feature space. Recent work has achieved striking predictive performance by using word embeddings or large language model activations to represent story stimuli, but at the cost of interpretability. These approaches risk explaining one black box in terms of another. Because these embeddings exhibit superposition—many concepts compressed into a limited set of shared directions—abstract notions such as time, space, and number become entangled. A voxel that appears to respond jointly to different concepts may reflect a genuinely shared neural code, or merely correlations inherent to the embedding. Until now, this ambiguity has been an unresolved confounding factor in semantic maps

reported in earlier works.

The Sparse Concept Encoding Model addresses this problem by first unpacking the embedding space. Using sparse dictionary learning, we factor dense word embeddings into a higher-dimensional basis of sparse “concept atoms”, each capturing a more coherent semantic direction. Encoding models are then refit in this new space, so that each voxel’s tuning is expressed as a sparse linear combination of concept atoms rather than a dense mixture of opaque coordinates. In this representation, each axis carries a clear conceptual meaning, allowing us to directly read out how time-, space-, and number-related atoms are expressed across cortex. Thus, we transform black-box features into a structured hypothesis space for understanding how abstract ideas are laid out in the brain.

1.2 Deep Brain Stimulation Amplifies Movement-Related Neural Modulation in Parkinson’s Disease

In Chapter 3, I pivot from distributed semantic representations in fMRI to a clinical setting where understanding neural dynamics carries immediate therapeutic stakes: deep brain stimulation (DBS) for Parkinson’s disease (PD). PD is a dynamical disease. Its motor symptoms arise not from a single static lesion, but from aberrant, time-varying activity patterns distributed across basal ganglia and sensorimotor cortex—patterns that are continuously shaped by movement, sensory input, medication state, and everyday context. Yet standard DBS is still delivered as an open-loop therapy: high-frequency stimulation runs continuously, and parameters are tuned largely by clinical intuition rather than by explicit models of how stimulation interacts with neural dynamics during behavior. Advancing DBS toward adaptive, principled neuromodulation requires something closer to what engineers would call *system identification*: measuring the system densely, in the regimes where it naturally operates, and learning the mapping between behavior, stimulation, and neural state.

Only recently has that ambition become realistic. Bidirectional neurostimulators can now deliver therapeutic stimulation while simultaneously recording local field potentials (LFPs) chronically in freely moving patients. At the same time, wrist-worn accelerometers and video-based pose estimation can capture behavior with a richness that the clinic cannot. Chapter 3 brings these tools together in a home-deployed multimodal platform that pairs an investigational Medtronic Summit RC+S device with bilateral wrist accelerometry and synchronized video. Two participants completed six recording days at home, performing a standardized UPDRS-derived motor battery under three monopolar stimulation amplitudes spanning under-stimulation, clinically preferred stimulation, and over-stimulation.

The encoding model framework mirrors the one in Chapter 2, but with different variables. Instead of asking how semantic features predict voxel responses, we ask how

behavioral measurements predict neural spectra recorded from the stimulation target and sensorimotor cortex. We treat brain–behavior coupling as an encoding problem: regularized linear models predict time-resolved neural log-power spectra from behavioral feature sets. This framing is deliberately modest—linear models, standard regularization—but it is exactly what makes the results interpretable enough to support mechanistic claims.

Three findings emerge. First, behavioral features robustly predict neural spectral activity across STN, precentral cortex, and postcentral cortex, showing that movement-related structure is reliably present in the signals sensed by the implanted system. Second, richer behavioral representations—especially one-hot task labels and the full watch-acceleration spectrogram—substantially outperform coarse summaries such as mean acceleration, highlighting how much information in neural structure is invisible to the conventional “rest versus movement” lens. Third, in the participant with stronger overall encoding, higher stimulation amplitudes increase prediction performance across the sensorimotor network by amplifying a distinctive movement-linked spectral signature, characterized by beta suppression alongside elevated gamma power, including stimulation-entrained gamma. In other words, stimulation does not merely push neural activity up or down; it changes the way movement-related modulation is written into the circuit’s rhythms, an effect that fits naturally with network-level accounts in which DBS alters pathological dynamics by changing patterns of information flow.

1.3 Summary and Motivation

Although these two projects operate at very different scales of the nervous system and draw on distinct data modalities—fMRI measurements of distributed semantic representations and LFPs from basal ganglia–cortical circuits during DBS—they are guided by the same philosophy. Both begin with high-volume, within-individual datasets collected in naturalistic or semi-naturalistic settings, where the dynamics of interest are actually expressed rather than averaged away. Both treat machine learning as a disciplined tool for linking rich measurements to latent structure, using explicit regularization and stringent held-out evaluation to emphasize generalization over post hoc significance reporting. And in both cases, the models are built to support explanation: their internal organization and fitted parameters can be read as concrete, testable hypotheses about brain function, serving as interpretable instruments for scientific insight rather than black boxes optimized for benchmark performance.

The techniques and computational resources that make these studies possible—large-scale optimization, pose estimation from video, word embeddings and sparse coding, would have been extremely difficult to deploy in practice even a decade ago. Yet in the rapidly evolving landscape of AI, they have already begun to feel like the tools of a previous generation. Despite this, it is clear that we are still just at the beginning of this sea change. AI is being woven into the entire scientific pipeline: proposing hypotheses,

designing experiments, controlling instruments, collecting and labeling data, and interpreting complex, multimodal datasets. It has already helped address problems that not long ago seemed intractable, such as large-scale protein structure prediction and certain classes of climate and molecular simulations. At the same time, there remain deep gaps in how we design algorithms, experiments, and institutions suited to this new regime of “AI for science.” Working through those gaps will be one of the central tasks of my generation of scientists and engineers.

“It was the best of times, it was the worst of times, it was the age of wisdom, it was the age of foolishness, it was the epoch of belief, it was the epoch of incredulity.” We are entering a moment that is at once exhilarating and perilous. Like all powerful technologies, the impact of machine learning tools depends on the social, scientific, and moral frameworks into which they are embedded. The same tools that can help us understand the brain and treat disease can also, if deployed carelessly, drive unemployment, dislocation, and social upheaval. They can be harnessed to construct new weapons and systems of surveillance or develop therapies and technologies that benefit humans and other living beings.

This thesis aspires to be a step along the latter path. It takes as its premise that machine learning can be a principled partner in scientific discovery. But it does not assume that such an outcome is guaranteed. What we choose to build, and how we choose to use it, remains, for now, in our hands.

CHAPTER 2

Disentangling Superpositions: Interpretable Brain Encoding Model with Sparse Concept Atoms

2.1 Abstract

Encoding models based on word embeddings or artificial neural network (ANN) features reliably predict brain responses to naturalistic stimuli but remain difficult to interpret. A central limitation is superposition—the entanglement of distinct semantic features along correlated directions in dense embeddings, which arises when latent features outnumber the embedding dimensions. This entanglement renders regression weights non-identifiable: different combinations of semantic directions can produce identical predicted brain activity, preventing principled interpretation of voxel selectivity. To overcome this, we introduce the Sparse Concept Encoding Model, which transforms dense embeddings into a higher-dimensional, sparse, and non-negative space of learned concept atoms. This transformation yields an axis-aligned semantic basis where each dimension corresponds to an interpretable concept, enabling direct readout of conceptual selectivity from voxel weights. When applied to fMRI data collected during story listening, our model matches the prediction performance of conventional dense models while substantially enhancing interpretability. It enables novel neuroscientific analyses such as disentangling overlapping cortical representations of time, space, and number, and revealing structured similarity among distributed conceptual maps. This framework offers a scalable and interpretable bridge between ANN-derived features and human conceptual representations in the brain.

2.2 Introduction

Artificial neural networks (ANNs) were originally inspired by the brain, yet progress in machine learning has far outpaced advances in understanding brain function. A

central bottleneck in neuroscience is data: while many large-scale datasets are available for training machine learning models, brain recordings remain scarce and costly. To overcome this limitation, neuroscientists increasingly use machine learning models to study how the brain represents information. A common framework for this approach is the **encoding model**, which predicts brain activity from quantitative stimulus features using linear regression (Huth et al., 2016; Leonard et al., 2024; Dupré la Tour et al., 2025; Nunez-Elizalde et al., 2019; Wu et al., 2006). Early encoding models relied on manually-defined features, such as labeled objects or actions in movies (Huth et al., 2012; Wehbe et al., 2014a; Nakai and Nishimoto, 2020). Recent encoding models instead leverage features derived from machine learning, such as word embeddings or activations from large language models (LLMs) (la Tour et al., 2021; Mathis et al., 2024; Oliver et al., 2024; Yamins and DiCarlo, 2016; Antonello et al., 2023; Jain and Huth, 2018; Goldstein et al., 2022; Li et al., 2023; Tang et al., 2023a; Kumar et al., 2023). When used as regressors to predict brain responses to naturalistic stimuli such as narrative stories, machine-learning-derived features consistently outperform manually-defined features. In particular, features from deeper neural networks achieve higher predictive accuracy (Antonello et al., 2023; Bonnasse-Gahot and Pallier, 2024; Schrimpf et al., 2020).

Despite their success in applications like speech decoding (Littlejohn et al., 2025; Tang et al., 2023b; Silva et al., 2024), ANN-based encoding models remain limited in their ability to advance fundamental neuroscience due to poor interpretability. Mapping ANN-derived representations onto brain activity often amounts to explaining one black box with another, yielding little insight into the underlying neural mechanisms. However, many core questions in neuroscience, such as where and how abstract concepts are represented in the brain, demand models with interpretable internal structure.

In this work, we address the interpretability challenges inherent to dense feature representations in encoding models. Previous studies using word embeddings or ANN-based features often assume that regression weights in these high-dimensional spaces are directly interpretable (Huth et al., 2016; Tang et al., 2023a; Wehbe et al., 2014b; Deniz et al., 2019). We show that this assumption is fundamentally flawed. Specifically, we demonstrate that *superposition*, the entanglement of multiple latent semantic features when their number exceeds the embedding dimensionality, introduces intrinsic ambiguity in model interpretation (Elhage et al., 2022; Bricken et al., 2023). When semantic features are superposed, regression weights become **non-identifiable**: distinct combinations of semantic features can produce the same predicted brain responses. This makes it impossible to reliably identify which specific concepts a voxel encodes, limiting the utility of dense models for neuroscientific inquiries. To overcome this limitation, we propose a new framework that preserves the predictive power of dense features while significantly enhancing model interpretability. Our method disentangles superposed representations into sparse, semantically-aligned concept dimensions, allowing a clearer link between model structure and brain activity. The framework generalizes naturally across a wide

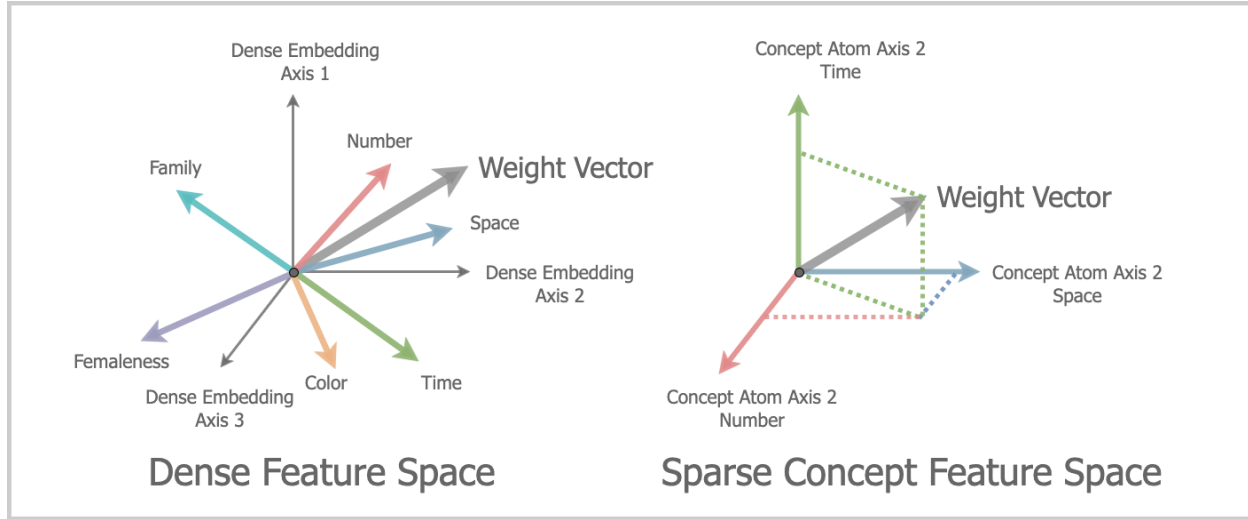


Figure 2.1: The Sparse Concept Encoding Model resolves superposition in dense feature spaces. **Left:** In dense embedding space, semantic concepts are represented by correlated linear combinations of basis directions. This superposition introduces ambiguity: a voxel’s weight vector may project positively onto multiple concept directions simply because those directions are themselves correlated. **Right:** The Sparse Concept Encoding Model recovers the underlying semantic directions and assigns each to an independent axis, transforming the dense space into an interpretable, axis-aligned representation. This enables voxel tuning to be directly read out along concept dimensions.

range of representations, such as LLM embeddings, and across recording modalities, such as electrocorticography (ECoG) and magnetoencephalography (MEG).

Main Contributions (1) We show that regression weights learned from dense features are inherently non-identifiable. Using a Bayesian analysis, we trace this ambiguity to the implicit prior induced by the geometry of dense embedding spaces. (2) We introduce the Sparse Concept Encoding Model, which transforms dense features into a higher-dimensional, sparse space composed of learned concept atoms. (3) We propose Vector Norm Reparameterization (VNR), a preprocessing method that enhances interpretability of learned concept atoms by expanding the relative spread of low-norm vectors while compressing the range of high-norm vectors. (4) We apply the Sparse Concept Encoding Model to a naturalistic fMRI dataset and demonstrate two neuroscientific analyses that are infeasible with dense models: (i) disentangling overlapping cortical representations of time, space, and number, and (ii) comparing concept representations across the cortex to reveal structured relationships among distributed conceptual maps.

2.3 Background

2.3.1 Sparse Coding

Sparse coding, also known as sparse dictionary learning, was originally proposed as a model for how the visual cortex efficiently represents natural scenes (Olshausen and Field, 1996, 1997). The central idea is that many natural signals, such as images, sounds, or words, can be expressed as linear combinations of a small number of latent components, often called *basis factors* or *atoms*. For example, the word "queen" can be decomposed into two latent factors: "royalty" and "femaleness". These features tend to occur independently: "royalty" appears in "king" and "castle", while "femaleness" is present in "daughter" and "mother". Since most words activate only a small number of such factors, the resulting representations are **sparse**.

Formally, each input $\mathbf{x}_j \in \mathbb{R}^p$ can be approximated by a linear combination of atoms from an overcomplete dictionary $\{\boldsymbol{\phi}_i\}_{i=1}^m \subset \mathbb{R}^p$, where $m > p$. The representation $\mathbf{x}_j = \sum_{i=1}^m a_{ij} \boldsymbol{\phi}_i$ uses a sparse coefficient vector $\mathbf{a}_j \in \mathbb{R}^m$ containing mostly zero entries. Overcompleteness ($m > p$) enables the dictionary to capture fine-grained structure in the data but renders the system underdetermined. A unique and interpretable solution is obtained by enforcing a regularized least-squares objective:

$$\min_{\{\mathbf{a}_j\}, \{\boldsymbol{\phi}_i\}} \sum_{j=1}^n \left\| \mathbf{x}_j - \sum_{i=1}^m a_{ij} \boldsymbol{\phi}_i \right\|_2^2 + \lambda \sum_{j=1}^n \|\mathbf{a}_j\|_1,$$

where the first term minimizes reconstruction error and the second imposes an ℓ_1 -penalty to promote sparsity. The hyperparameter λ balances fidelity and sparsity. From a Bayesian perspective, the ℓ_1 term corresponds to the negative log-prior of a Laplace distribution on $\mathbf{a}_j$, yielding a Maximum A Posteriori (MAP) estimate of the codes and dictionary.

2.3.2 Sparse Coding For Interpreting Word Embeddings and Deep Neural Networks

Word embeddings are known to encode semantic features as linear directions in activation spaces. The canonical example by Mikolov et al. (2013), "king" - "man" + "woman" $\approx$ "queen", showed that vector arithmetic on word embeddings captures analogies. This suggests that certain linear directions correspond to interpretable semantic concepts ("royalty" and "femaleness"). Arora et al. (2016) provided a theoretical explanation for the linear structure through a *random-walk-on-discourses* model, where word vectors are generated from mixtures of latent discourse vectors. Arora et al. (2018) further argued that sparse coding provides a natural method to recover these latent discourse directions from dense embeddings (Faruqui et al., 2015). Subsequent studies applying sparse coding to word embeddings have indeed revealed dictionary atoms that correspond to interpretable

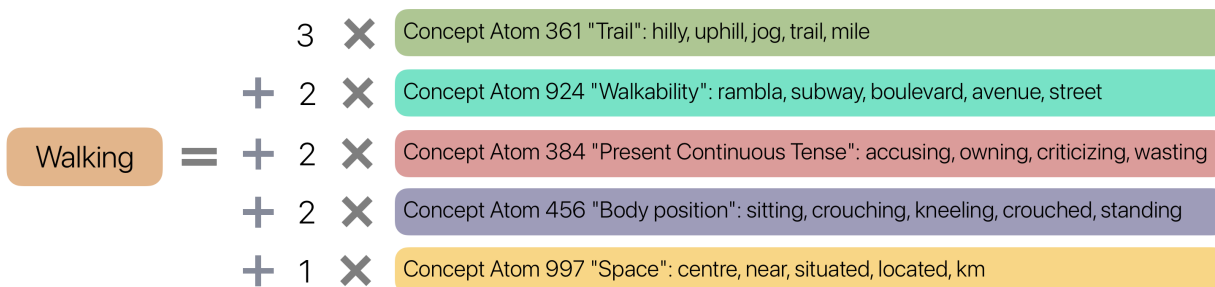


Figure 2.2: Sparse dictionary learning decomposes each word into a sparse, non-negative linear combination of concept atoms. This example shows the sparse decomposition of the word “walking” using the learned concept atom dictionary. Each contributing atom is shown with its index, a descriptive label (chosen heuristically by the authors), and its top 5 activating stimulus words from the story transcripts.

semantic and syntactic features (Arora et al., 2018; Balogh et al., 2020; Zhang et al., 2021). Recently, sparse coding has been adopted in mechanistic interpretability research on deep neural networks (DNNs), which aims to decompose models into human-interpretable components (Räuker et al., 2023; Sharkey et al., 2025). Early efforts focused on analyzing individual neurons, under the assumption that each neuron corresponds to a single interpretable feature (Karpathy et al., 2015; Olah et al., 2017; Bau et al., 2017). However, empirical studies revealed that many neurons respond to multiple, often unrelated features, a phenomenon known as *polysemanticity* (Olah et al., 2017). To explain this, researchers proposed the *superposition hypothesis* (Elhage et al., 2022; Bricken et al., 2023): when the number of latent features exceeds the number of available neurons, the network encodes each feature as a distinct linear combination of neurons. Superposition thus emerges naturally as an efficient representational strategy under data sparsity and limited capacity. Inspired by earlier work on word embeddings, recent studies have applied sparse coding to dense DNN activations to recover interpretable basis factors (Cunningham et al., 2023; Yun et al., 2023; Gurnee et al., 2023). In particular, sparse autoencoders trained on large language model (LLM) activations have shown that a few hundred neurons can represent thousands of distinct, interpretable features in superposition (Bricken et al., 2023).

2.4 Theory

2.4.1 Limitations of Dense Encoding Models

The seminal work of Huth et al. (2016) used word embeddings to represent narrative stimuli and revealed widespread conceptual representations across the human cortex. This approach catalyzed a growing body of research using machine-learning-derived features

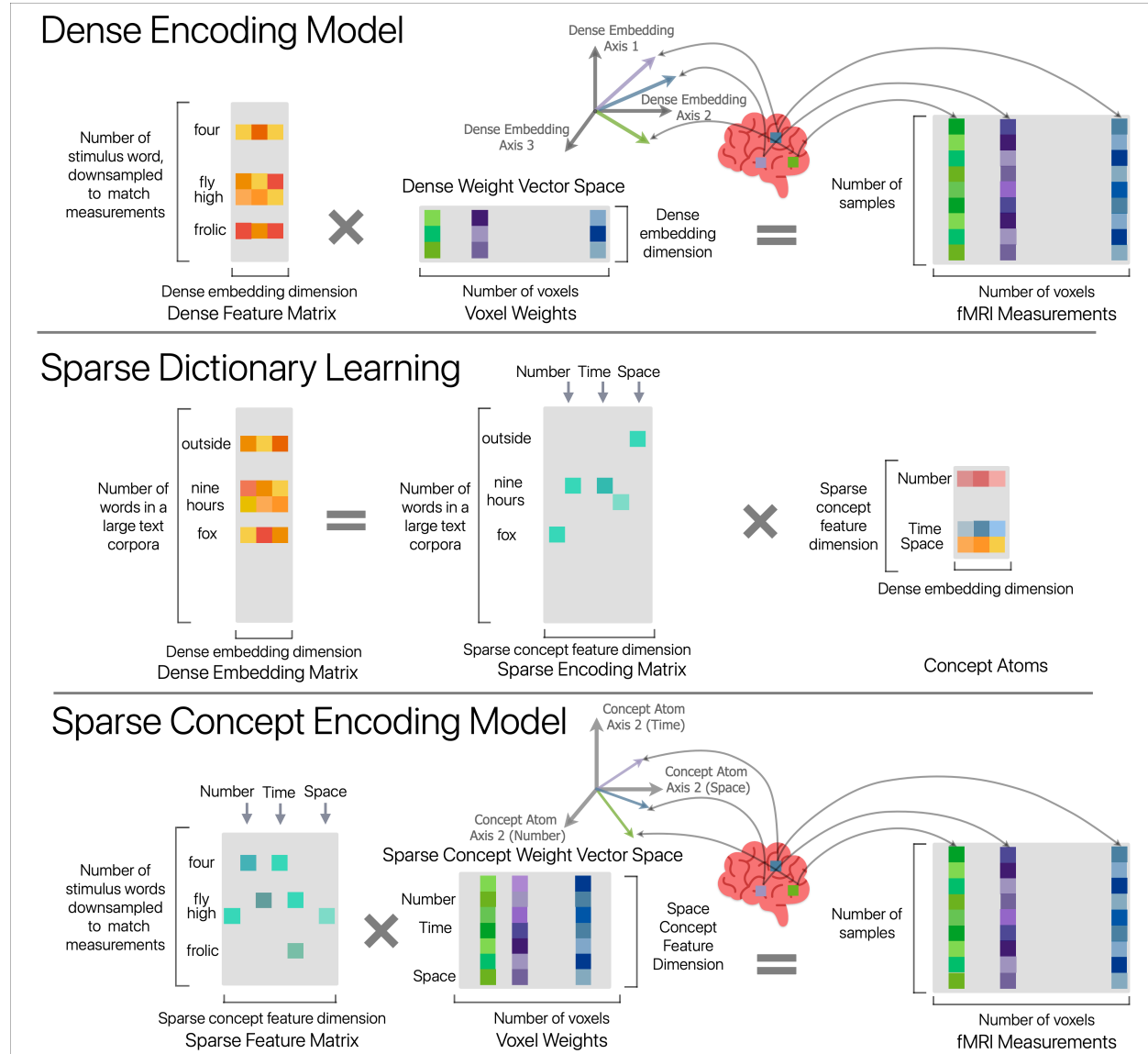


Figure 2.3: The Sparse Concept Encoding Model produces interpretable voxel weights by replacing dense features with learned concept atoms. Top: *Dense Encoding Model*: Stimuli are represented using dense word embeddings, resulting in voxel weights that lie in an entangled feature space where semantic directions overlap. **Middle: *Sparse Dictionary Learning*:** Dense embeddings are transformed into a higher-dimensional space of sparse concept atoms, each defining a distinct, interpretable semantic axis. **Bottom: *Sparse Concept Encoding Model*:** Stimuli are re-encoded as sparse combinations of concept atoms. The resulting voxel weights are defined with respect to disentangled concept-atom axes, enabling direct and interpretable readout of semantic selectivity.

to predict brain activity. Most studies employ regularized linear regression models that predict brain activity from quantitative stimulus features (Dupré la Tour et al., 2025; Nunez-Elizalde et al., 2019; Mathis et al., 2024). A separate regression model is fit for each voxel, a volumetric unit of brain tissue typically measuring approximately $3 \times 3 \times 4$ millimeters, yielding a weight vector $w_j \in \mathbb{R}^p$ that describes the voxel’s feature selectivity:

$$\min_{w_j} |y_j - Xw_j|_2^2 + \lambda_j |w_j|_2^2, \quad (2.1)$$

where y_j is the response vector for voxel j .

However, interpreting regression weights from dense encoding models presents a fundamental ambiguity. In dense embeddings (Fig. 2.1, left), the coordinate axes are not inherently interpretable; semantic concepts are represented by correlated linear combinations of basis directions. This entangled structure complicates interpretation, as the relationship between voxel weights and semantic concepts becomes opaque. Conventional analyses often rely on heuristic assumptions: for example, inferring that a voxel is selective for a concept if its weight vector projects positively onto that concept’s direction. But such inferences are unreliable when concept directions are correlated. A voxel aligned with *number*, for instance, may appear to also respond to *time* simply because the two concept directions are positively correlated. More problematically, many concept pairs, such as *femaleness* and *space* in the diagram, may be strongly anti-correlated, making it unclear whether a voxel is activated by one, suppressed by the other, or both. These ambiguities undermine principled interpretation: a positive projection does not necessarily imply true selectivity to a concept; observed overlaps in cortical maps may reflect either genuine multi-concept selectivity or spurious effects induced by embedding geometry. The root of this ambiguity is formalized in the next section. In contrast, an ideal feature space would have an interpretable basis in which each axis corresponds to a distinct semantic concept (Fig. 2.1, right). In such a space, a voxel’s tuning could be directly read from its projections along concept axes. This axis-aligned structure is the interpretability goal we aim to achieve with the Sparse Concept Encoding Model.

2.4.2 Theorem (Non-identifiability in Dense Embedding Spaces)

Let $\{\mathbf{f}_i\}_{i=1}^k \subset \mathbb{R}^d$ be unit-norm concept atoms, and define the Gram matrix:

$$G = [G_{ij}]_{i,j=1}^k, \quad G_{ij} = \langle \mathbf{f}_i, \mathbf{f}_j \rangle.$$

Suppose a voxel’s weight vector is tuned to a linear combination of these atoms:

$$\mathbf{w} = \sum_{i=1}^k \alpha_i \mathbf{f}_i, \quad \boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_k)^\top.$$

Taking inner products with each $\mathbf{f}_j$ yields the observed projections:

$$\begin{pmatrix} \langle \mathbf{w}, \mathbf{f}_1 \rangle \\ \vdots \\ \langle \mathbf{w}, \mathbf{f}_k \rangle \end{pmatrix} = G\boldsymbol{\alpha}.$$

If G is invertible, then $\boldsymbol{\alpha}$ can be uniquely recovered. However, if G is singular, the system is non-identifiable: multiple combinations of concept atoms can yield the same weight vector. In the ideal case where all $\mathbf{f}_j$ are orthogonal, $G = I$ and $\boldsymbol{\alpha}$ can be directly recovered.

2.4.3 The Sparse Concept Encoding Model

To circumvent the identifiability limitations of dense embeddings, we aim to recover the latent semantic directions and assign each its own coordinate, effectively lifting the dense feature space into an interpretable, axis-aligned representation (Fig. 2.1, right). This is achieved using non-negative sparse dictionary learning on a large text corpus, yielding an overcomplete set of semantic axes, which we refer to as **concept atoms**. Specifically, we expand the original d -dimensional embedding space into an m -dimensional space ($m > d$), where each axis corresponds to a distinct concept atom (Fig. 2.3, middle). Following dictionary learning, each stimulus vector in the dense space is re-encoded as a sparse, non-negative linear combination of concept atoms (see Fig. 2.2 for an example). This transformed representation replaces the original dense features in the encoding model. In the resulting sparse feature space, voxel weights are directly interpretable: each dimension aligns with a distinct semantic concept, allowing principled readout of conceptual selectivity from the weights along corresponding concept-atom axes (Fig. 2.3, bottom). This structure enables clean disentanglement of voxels jointly selective for multiple concepts. The non-negativity constraint further sharpens interpretability by ensuring that positive weights indicate activation and negative weights indicate suppression. Together, these properties enable disambiguated interpretation of conceptual selectivity across the cortex.

Both the Sparse and Dense Encoding Models can be interpreted under a Bayesian regression framework (Nunez-Elizalde et al., 2019). From this view, the Dense Model corresponds to the Sparse Model with a non-spherical Gaussian prior whose covariance captures correlations among concept directions in the dense space.

2.4.4 Bayesian Interpretation of Sparse and Dense Encoding Models

Both the Sparse Concept Encoding Model and the Dense Encoding Model admit a Bayesian linear regression interpretation (Nunez-Elizalde et al., 2019). This view makes explicit the prior assumptions each model imposes on voxel tuning over semantic features.

Sparse Concept Encoding Model. Each stimulus is represented by a sparse, non-negative activation over m learned concept atoms, yielding a design matrix $Z \in \mathbb{R}^{T \times m}$. Voxel responses are fit with ridge regression

$$\hat{\beta} = \arg \min_{\beta \in \mathbb{R}^m} \|\mathbf{y} - Z\beta\|_2^2 + \lambda \|\beta\|_2^2,$$

which corresponds to MAP inference under an isotropic Gaussian prior

$$\beta \sim \mathcal{N}(0, \lambda^{-1}I).$$

This prior assumes independent contributions of each concept atom with equal variance.

Dense Encoding Model. Let the dense feature matrix be factorized as $X_{\text{dense}} = Z\Phi$, where $\Phi \in \mathbb{R}^{m \times d}$ maps dense coordinates to the m concept atoms (rows of Φ are atom vectors), and fit ridge regression in dense space:

$$\hat{\mathbf{w}} = \arg \min_{\mathbf{w} \in \mathbb{R}^d} \|\mathbf{y} - Z\Phi\mathbf{w}\|_2^2 + \lambda \|\mathbf{w}\|_2^2.$$

Define the implied sparse-space weights $\beta := \Phi\mathbf{w} \in \mathbb{R}^m$. As shown below, dense ridge is equivalent to a sparse-space regression with a structured (generally degenerate) Gaussian prior

$$\beta \sim \mathcal{N}(0, \lambda^{-1}\Phi\Phi^\top).$$

When $\Phi\Phi^\top$ is singular, this is a degenerate Gaussian supported on $\mathcal{C}(\Phi)$; its MAP penalty is $\beta^\top (\Phi\Phi^\top)^+ \beta$. Intuitively, atoms with similar directions (large dot product) will be biased to have coupled weights under the prior. This coupling can improve prediction when related concepts truly co-activate, but it complicates interpretation: voxel weights may reflect embedding correlations rather than true semantic selectivity.

2.4.4.1 Proof

Start from dense ridge:

$$\hat{\mathbf{w}} = \arg \min_{\mathbf{w} \in \mathbb{R}^d} \|\mathbf{y} - Z\Phi\mathbf{w}\|_2^2 + \lambda \|\mathbf{w}\|_2^2.$$

Let $\beta := \Phi\mathbf{w}$. Then $\|\mathbf{y} - Z\Phi\mathbf{w}\|_2^2 = \|\mathbf{y} - Z\beta\|_2^2$, and feasible β satisfy $\beta \in \mathcal{C}(\Phi)$.

Case A: $\text{rank}(\Phi) = d$ (injective). For each $\beta \in \mathcal{C}(\Phi)$ there is a unique $\mathbf{w}$ such that $\Phi\mathbf{w} = \beta$, given by

$$\mathbf{w} = (\Phi^\top \Phi)^{-1} \Phi^\top \beta.$$

Substituting into the objective yields

$$\hat{\beta} = \arg \min_{\beta \in \mathcal{C}(\Phi)} \|\mathbf{y} - Z\beta\|_2^2 + \lambda \beta^\top (\Phi\Phi^\top)^+ \beta.$$

If $m = d$ and Φ is invertible, then $\mathcal{C}(\Phi) = \mathbb{R}^m$ and $(\Phi\Phi^\top)^+ = (\Phi\Phi^\top)^{-1}$.

Case B: $\text{rank}(\Phi) < d$ (non-injective). For fixed $\beta \in \mathcal{C}(\Phi)$, the constraint $\Phi\mathbf{w} = \beta$ admits infinitely many solutions. Moreover, all $\mathbf{w}$ satisfying $\Phi\mathbf{w} = \beta$ yield the same prediction $Z\Phi\mathbf{w} = Z\beta$, so the data-fit term is constant over this solution set. Dense ridge therefore selects the solution with minimal norm,

$$\mathbf{w}_{\min} = \arg \min_{\mathbf{w}: \Phi\mathbf{w}=\beta} \|\mathbf{w}\|_2^2,$$

which is given by the Moore–Penrose pseudoinverse,

$$\mathbf{w}_{\min} = \Phi^\dagger \beta, \quad \Phi^\dagger = (\Phi^\top \Phi)^+ \Phi^\top.$$

Substituting $\mathbf{w}_{\min}$ into the regularizer gives

$$\min_{\mathbf{w}: \Phi\mathbf{w}=\beta} \|\mathbf{w}\|_2^2 = \|\Phi^\dagger \beta\|_2^2 = \beta^\top (\Phi\Phi^\top)^+ \beta,$$

and thus recovers the same sparse-space objective as in Case A:

$$\hat{\beta} = \arg \min_{\beta \in \mathcal{C}(\Phi)} \|\mathbf{y} - Z\beta\|_2^2 + \lambda \beta^\top (\Phi\Phi^\top)^+ \beta.$$

2.5 Methods

2.5.1 Experimental Procedure

The dataset used in this study was originally collected for a previously published experiment (Huth et al., 2016). All procedures related to participant recruitment, scanning parameters, and data acquisition were described in that work. Briefly, each participant completed eleven fMRI scans while listening to narrative audio stories from *The Moth Radio Hour*, in which individuals recount autobiographical experiences before a live audience. Each scan featured a single story lasting 10–15 minutes, with 10 seconds of silence preceding and following the audio.

2.5.2 Dataset Structure and Test/Train Split

The full dataset comprised eleven stories per participant. Ten stories were used for training the encoding models, yielding approximately 125 minutes of data per subject. One remaining story was held out for testing and presented twice to each participant. To improve the signal-to-noise ratio in the test data, BOLD responses across the two repetitions were averaged. Final model evaluations were conducted on this averaged 10-minute test set.

2.5.3 fMRI Preprocessing

Functional MRI data were preprocessed using the same pipeline as described in a previously published study (Meschke et al., 2023). The resulting fMRI response matrices consisted of approximately 60,000–80,000 voxels and 3,717 time points per subject. All analyses were conducted in each subject’s native voxel space, with no spatial smoothing applied.

2.5.4 Feature Construction

Low-Level Acoustic Features. To control for correlations between semantic content and acoustic structure, we included 41 nuisance regressors following (Deniz et al., 2019):

- 39 phoneme-rate features reflecting the frequency of each English phoneme over time
- A single feature representing total phoneme rate
- A single feature representing word rate

Semantic Features. Each word in the stimulus transcript was transformed into two parallel semantic representations:

- 300-dimensional dense GloVe embeddings (Pennington et al., 2014)
- 1,000-dimensional sparse concept embeddings derived via dictionary learning from GloVe

All features were resampled to match the fMRI acquisition rate using a 3-lobe Lanczos low-pass filter with a cutoff of 0.249 Hz (the Nyquist frequency). This resulted in feature matrices of size $3,717 \times 300$ for the Dense Encoding Model and $3,717 \times 1,000$ for the Sparse Concept Encoding Model.

To account for hemodynamic lag, a finite impulse response (FIR) model was applied to the design matrix. Each feature time series was delayed by 1, 2, 3, and 4 TRs (corresponding to 2, 4, 6, and 8 seconds) and concatenated, allowing the model to estimate a linear temporal kernel per feature (Dupré la Tour et al., 2025).

2.5.5 Feature Normalization

All stimulus features were demeaned over time within each scan run, except for the 300-dimensional GloVe embeddings. These embeddings were left unchanged to maintain a consistent origin across runs. Because cosine similarity depends on vector direction relative to a fixed origin, shifting the origin across runs would undermine interpretability. In contrast, sparse concept features and low-level acoustic features were demeaned, as

each feature dimension carries a specific, interpretable meaning. For these features, mean-centering does not affect the interpretation of voxel weights.

2.5.6 Model Estimation

Voxelwise encoding models were estimated using banded ridge regression (Dupré la Tour et al., 2022), which applies separate regularization to the semantic and low-level feature subspaces. For each voxel, two regularization hyperparameters ($\lambda_{\text{semantic}}$ and $\lambda_{\text{low-level}}$) were optimized using leave-one-run-out cross-validation on the training set.

The procedure was as follows:

- The 10 training runs were split into 9 training and 1 validation run, iterated over all folds.
- For each fold, 20 logarithmically spaced values (from 10^1 to 10^{20}) were tested for each hyperparameter.
- Prediction accuracy was computed on the held-out run.
- The hyperparameter pair yielding the highest average prediction accuracy across folds was selected separately for each voxel.

Model fitting was implemented using the Himalaya Python library (Dupré la Tour et al., 2022)¹, which enables efficient banded ridge regression on both CPUs and GPUs. Training required approximately 30–60 minutes per subject on a single NVIDIA RTX A6000 GPU.

2.5.7 Model Evaluation

Final models were evaluated on the held-out test story using the average blood-oxygen-level-dependent (BOLD) responses from two repeated presentations. Prediction accuracy was computed as the coefficient of determination (R^2) between predicted and observed fMRI responses, independently for each voxel.

2.5.8 Postprocessing of Semantic Weights

Model weights were first averaged across the five FIR delays to yield a single value per feature per voxel. To remove differences in overall scale introduced by voxel-specific ridge regularization, each semantic weight vector was normalized to unit length. Vectors were then scaled by the voxel’s cross-validated prediction accuracy, allowing downstream analyses to reflect both tuning direction and model reliability. Negative accuracies were clipped to zero before scaling.

¹<https://github.com/gallantlab/himalaya>

2.5.9 Top 50 Sparse Concept Atoms and Associated Words

We identified the 50 concept atoms most strongly expressed in the Sparse Concept Encoding Model, based on their average activation across voxels. During analysis, we found that a small number of atoms exhibited high activation correlations with many others in the sparse feature matrix. This redundancy complicates interpretation by making it difficult to isolate distinct conceptual contributions to neural responses. To address this, we excluded highly redundant atoms from all subsequent analyses. Specifically, we computed the pairwise Pearson correlation matrix across all 1,000 sparse feature dimensions and calculated the average absolute correlation of each atom with all others. Atoms whose average correlation exceeded 70% of the maximum observed value were removed. This thresholding procedure yielded a decorrelated subset of concept atoms better suited for interpreting voxelwise model weights. From the remaining atoms, we computed the mean absolute activation across voxels using the postprocessed model weights. The 50 atoms with the highest average activations were selected, as they were most strongly engaged across the brain. To interpret the semantic content of each atom, we identified the most strongly associated words from the training stories. Only words with an activation value greater than 0.2 for a given atom were retained. For each selected atom, the 20 highest-ranked words (sorted by activation strength) are listed in Table 2.1, along with the atom's original index from the full 1,000-dimensional dictionary.

2.5.10 Semantic Category Selection for Time, Number, and Space

To identify concept atoms associated with the semantic domains of time, number, and space, we began by defining representative seed word lists for each category. Time-related words included *second, hour, day, week, month, year, today, and ago*; number-related terms included *one, two, three, ten, fifteen, hundred, and thousand*; and spatial expressions included *near, far, distance, mile, outside, map, and direction*.

For each category, we ranked all 1,000 concept atoms by their average activation across the corresponding seed words and manually examined the top five. We retained atoms whose top associated words were semantically aligned with the target category, using inclusive criteria to ensure coverage of relevant conceptual variations. Table 2.2 summarizes the selected concept atoms along with their highest-activation words. These concept atoms capture distinct subtypes of meaning within each semantic domain. For example, factor 81 emphasizes time points typically used in “*at*” contexts (e.g., *at night, at 9 a.m.*); factor 798 encodes temporal durations often found in “*for*” constructions (e.g., *for days, for weeks*); and factor 880 reflects temporal eras or historical references (e.g., *seventies, grunge, persisted*). For the main analysis in Section 2.6.5, we selected one representative concept atom from each semantic category—factor 81 (time), factor 505 (number), and factor 997 (space).

2.5.11 Software Implementation

All analyses were implemented in Python using a custom framework, `sparseconcept` which is openly available on GitHub.² The analysis pipeline relies on standard scientific Python libraries including `numpy`, `scipy`, `matplotlib`, `scikit-learn`, `statsmodels` and `pycortex`, and uses the `himalaya` package with a PyTorch backend for efficient voxelwise model fitting.

2.6 Results

2.6.1 Voxelwise Modeling and Feature Construction

We evaluated the Sparse Concept Encoding Model on a publicly available fMRI dataset in which participants listened to naturalistic autobiographical audio stories (Huth et al., 2016). Our analysis follows the voxelwise encoding model framework, emphasizing high-resolution, within-subject modeling in which model training and evaluation are performed on separate data from the same individual (Dupré la Tour et al., 2025). We present cortical maps for two representative participants and report group-level statistics across all seven participants.

To construct the dense feature space, we used pre-trained 300-dimensional GloVe word embeddings (Pennington et al., 2014). The corresponding sparse feature space was obtained using non-negative sparse dictionary learning, following the procedure of Zhang et al. (Zhang et al., 2021). Using their public implementation, we learned a 1,000-dimensional overcomplete dictionary of concept atoms from the GloVe vectors. For each atom, we identified the top-activating words from the story transcripts to aid interpretability. Figure 2.3 shows illustrative examples, and a broader list of 50 representative concept atoms is provided in Table 2.1 below.

A key question is whether the dense feature space already provides sufficient semantic separation for voxel weights to be interpreted directly. If distinct concepts were represented along nearly orthogonal directions, sparse transformation would be unnecessary. To evaluate this, we examined the pairwise cosine similarities among the 1,000 learned concept atoms in the original 300-dimensional GloVe space. For reference, 1,000 random unit vectors in a 300-dimensional space are expected to be nearly orthogonal: their pairwise cosine similarities have mean $\mu = 0$, standard deviation $\sigma = 0.058$, and only 0.03% of pairs exceed 0.2 in magnitude. In contrast, the concept atoms exhibited substantial overlap, with mean pairwise similarity $\mu = 0.01$, standard deviation $\sigma = 0.075$, and 2.18% of pairs exceeding 0.2 (Fig. 2.4a). This structured correlation pattern indicates that the dense feature space represents concepts along moderately correlated directions, reaffirming the need for a sparse, axis-aligned representation.

²<https://github.com/alicialitrtwe/sparseconcept>

Table 2.1: Fifty concept atoms with the highest average cortical activation, listed with their top associated words.

Atom index	Top associated words
505	eleven, sixteen, nineteen, fifteen, seventeen, twenty, ten, eighteen, thirty, sixty
562	usually, often, also, normally, may, actually, always, might, definitely, sometimes
254	jumped, ran, came, pulled, blew, proceeded, took, stood, fell, stopped
115	grandparents, sister, dad, cousins, brother, siblings, mom, grandfather, mother, stepmother
261	learn, want, able, take, join, to, need, begin, wish, go
585	call, telephone, appointment, phone, us, anytime, ask, queries, answering, number
670	business, personal, home, city, family, school, music, game, woman, life
369	although, though, however, despite, but, yet, anyway, surprisingly, somewhat, far
549	is, becomes, an, itself, a, exists, another, comes, represents, was
301	cars, dolls, sticks, boxes, pillows, shirts, cans, beds, are, cards
218	herself, queen, goddess, her, waitress, she, sister, mother, woman, where
672	died, since, graduated, moved, dated, began, until, lived, existed, was
304	easy, easier, quick, simple, instantly, allows, handy, lets, solution, fast
456	standing, sitting, stood, sat, sit, crouch, aisle, stands, lying, behind
798	months, hours, days, weeks, ago, years, minutes, hour, month, year
719	cow, bird, hat, cat, tiny, bag, ball, pink, baby, tree
893	pulls, sees, leans, asks, grabs, takes, waits, thinks, goes, gets
81	morning, afternoon, evening, weekends, monday, tuesday, lunch, night, hour, day
626	guy, sailor, boy, woman, waitress, man, doctor, stranger, mister, kid
335	cry, heart, eyes, soul, breath, lover, alive, dream, cruel, smile
172	acknowledged, insisted, suggested, explained, said, admitted, saying, recounted, claimed, told
707	thanks, hello, sorry, hey, bye, hi, dear, thank, yeah, ya
530	am, wrote, by, at, on, subject, anonymous, views, writes, abuse
184	discuss, topic, explored, similarities, about, examining, issues, examine, subject, dealing
148	toward, towards, forward, fro, slowly, direction, moving, into, back, headed
486	steve, dave, andy, tim, rob, kevin, matt, bob, todd, nick
733	from, this, notes, ago
244	insanely, remarkably, incredibly, terribly, extremely, perennially, oddly, horribly, surprisingly, quite
494	gon, outta, ya, gotta, lotta, em, wanna, gonna, fo, shit
17	page, here, visit, details, check, listed, list, information, above, see
997	near, close, distance, airport, center, miles, opposite, minutes, village, mile
385	um, o, em, eh, e, legal, ate, nova, do
750	many, other, several, common, similar, different, such, particularly, most, certain
321	awesome, great, amazing, lovely, nice, beautiful, gorgeous, brilliant, definitely, neat
384	talking, doing, enjoying, pushing, jumping, stopping, taking, chasing, throwing, keeping
766	everywhere, anywhere, happening, in, within, exist, occurs, occurred, exists, somewhere
782	brother, himself, father, son, his, grandfather, marquis, him, minister, afterwards
640	families, residents, people, adults, students, women, children, relatives, men, fellow
890	director, vice, professor, consultant, assistant, manager, says, said, joined, counselor
730	teachers, school, teacher, kindergarten, classroom, grades, grade, education, students, gifted
610	drank, counseled, listened, cared, worshipped, hated, travelled, liked, ate, lived
867	hands, finger, arms, arm, cheek, shoulders, thumb, rubbing, squeezes, ears
8	caused, cause, due, occurs, result, serious, suffered, failure, loss, problems
119	shorts, shirts, jeans, shirt, pants, jacket, cropped, vest, lauren, casual
264	bedroom, upstairs, room, downstairs, bathroom, beds, lobby, floor, closet, living
886	stabbed, stabbing, victim, fled, suspicious, blaze, killed, raped, man, missing
970	our, proud, committed, recognize, mission, ourselves, deeply, understands, tradition, honor
853	thin, tricky, slightly, tight, somewhat, bit, awkward, too, notoriously, low
700	alabama, virginia, arizona, florida, california, georgia, texas, vermont, washington, wyoming
842	event, held, attended, events, conference, booth, invited, host, weekend, fair

2.6.2 Improving Embedding Geometry with Vector Norm Reparameterization

A known property of dense word embeddings is that high-frequency words tend to lie closer to the origin, likely because their vectors represent averages over more diverse contextual usages (Schakel and Wilson, 2015; Arora et al., 2016). We observe that sparse coding often captures many high-frequency words with different semantic meanings within the same dictionary atom. To improve separability among high-frequency words, we developed a preprocessing technique called Vector Norm Reparameterization (VNR), which differentially rescales each word vector according to its norm, expanding the relative spread of vectors near the origin while compressing those among high-norm vectors. Quantitative comparisons with standard preprocessing methods showing that VNR improves WordNet purity, reconstruction error, and sparsity are included in Section 2.8 (Supplementary Materials) section. Based on these results, we adopted VNR for all subsequent analyses.

2.6.3 Comparing Predictive Performance of Dense and Sparse Models

To compare the predictive performance and interpretability of the Dense Encoding Model and the Sparse Concept Encoding Model, we fit voxelwise encoding models using two types of semantic features: 300-dimensional dense embeddings for the Dense Model and 1,000-dimensional sparse embeddings for the Sparse Model. Both models were trained using regularized linear regression with cross-validation and evaluated on a held-out test story using the coefficient of determination (R^2) as the performance metric. Statistical significance was assessed at the voxel level using 1,000 block permutations with false discovery rate (FDR) correction at $p < 0.05$ Section 2.8 (Supplementary Materials). To visualize prediction performance, we plotted voxelwise R^2 values onto cortical flatmaps (Gao et al., 2015). The Sparse Model achieved strong prediction accuracy across temporal, parietal, and prefrontal cortices in both participants (Fig. 2.4b–c). To directly compare models, we plotted voxelwise R^2 scores across all significantly predicted voxels (Fig. 2.4d–e). Across subjects, the mean difference in prediction performance between the transformed and original models was 0.00047 ± 0.00071 , with no statistically significant difference detected (paired $t(6) = -1.77$, $p = 0.13$). These results indicate that the Sparse Concept Encoding Model achieves comparable prediction accuracy to that of the Dense Encoding Model while substantially improving interpretability.

2.6.4 Concept-Level Cortical Maps from the Sparse Concept Encoding Model

Prior work has shown that abstract concepts are represented in the brain through widespread, spatially overlapping cortical patterns rather than anatomically localized

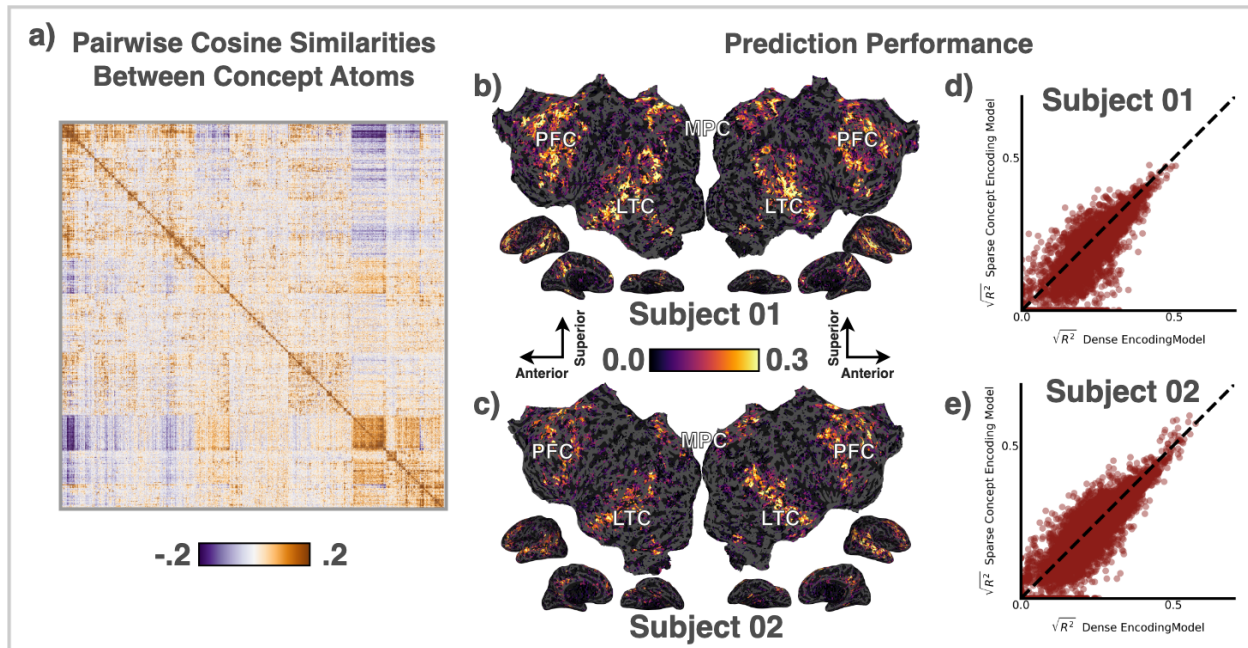


Figure 2.4: The Sparse Concept Encoding Model preserves prediction accuracy while improving interpretability. (a) Pairwise cosine similarities between all 1,000 concept atoms in the dense embedding space reveal structured correlations, with many atom pairs showing moderate cosine similarity. (b-c) Cortical maps of prediction accuracy ($\sqrt{R^2}$) for the Sparse Concept Encoding Model in Subject 01 and Subject 02, showing strong performance across lateral temporal (LTC), medial parietal (MPC), and prefrontal (PFC) regions. (d-e) Scatter plots comparing prediction accuracy between the Sparse and Dense Encoding Models across all significantly predicted voxels.

regions (Huth et al., 2012, 2016). However, dense encoding models lack a principled way to disentangle the cortical representations of individual concepts. The Sparse Concept Encoding Model addresses this by representing each concept as a distinct axis in a sparse feature space. These axes, or concept atoms, correspond to interpretable semantic dimensions, which can be identified by inspecting the top-activating stimulus words (see Table 2.1). Many atoms align with abstract categories such as “body position” or “oppositional conjunctions.” In this and Section 2.6.6, we focused on the 20 atoms with the highest average cortical activation, as they provide the most robust signal for analyzing concept-level representations. To assess whether the Sparse Concept Encoding Model yields consistent and interpretable cortical maps for individual semantic concepts, we examined two illustrative concept atoms, indices 456 and 115, drawn from the top 20 atoms (Fig. 2.5). These atoms are interpretable as representing the concepts of “body

position” and “family”, respectively. For each atom, we listed the most strongly activating stimulus words and, for two participants, visualized voxelwise regression weights along the corresponding concept-atom axis on cortical flatmaps. The resulting maps revealed bilateral, distributed patterns of selectivity that are consistent across individuals. As a baseline, we compared these maps to those derived from the Dense Encoding Model. Since the dense model lacks a principled method for generating concept-level maps, prior work has relied on heuristic strategies. Common approaches include: (1) assigning each voxel to the concept whose embedding vector has the highest cosine similarity to its weight vector, which can underestimate mixed selectivity; and (2) projecting voxel weights onto individual semantic directions, which can overestimate selectivity due to correlations in dense space. In this work, we implemented a hybrid heuristic: labeling a voxel as selective for a concept if that concept ranked among its top-10 cosine similarity matches. Maps produced by the dense model were less spatially coherent, less bilaterally symmetric, and more variable across individuals (2.8, Supplementary Materials).

To quantify inter-subject consistency across models, we projected each participant’s cortical maps onto the common fsaverage surface and computed pairwise cosine similarities across all seven individuals. Averaged across the top 20 concept atoms, the Sparse Concept Encoding Model yielded substantially higher inter-subject consistency (mean $\pm$ SD : 0.26 ± 0.04) compared to the Dense Encoding Model (0.09 ± 0.04). These results indicate that the sparse model produces semantic representations that are clearer, more disentangled, and more consistent across subjects.

2.6.5 Case Study: Disentangling Time, Space, and Number Representations

A longstanding hypothesis in cognitive neuroscience posits that the brain encodes time, space, and number using a shared magnitude system (Dehaene, 1997; Walsh, 2003; Dehaene and Brannon, 2010). While this theory is supported by robust behavioral and psychophysical evidence, neuroimaging results have been inconsistent (Cona et al., 2024; Simsek et al., 2025). The Sparse Concept Encoding Model enables a novel test of this hypothesis by assigning time, space, and number independent axes in a shared semantic space, allowing direct comparison of their cortical representations. To evaluate this, we identified the concept atoms most associated with each domain based on their top-activating stimulus words (Fig. 2.6). To normalize for variation in activation magnitude across concept atoms, we selected the top 5,000 voxels with the highest positive regression weights for each cortical map. The resulting composite maps revealed localized, bilateral overlap among the three conceptual domains, indicated in white. These patterns were consistent across individuals. Notably, regions of three-way overlap included the bilateral intraparietal sulcus (IPS) and bilateral inferior frontal gyrus (IFG), areas previously implicated in shared magnitude processing (Cona et al., 2024; Simsek et al., 2025). To

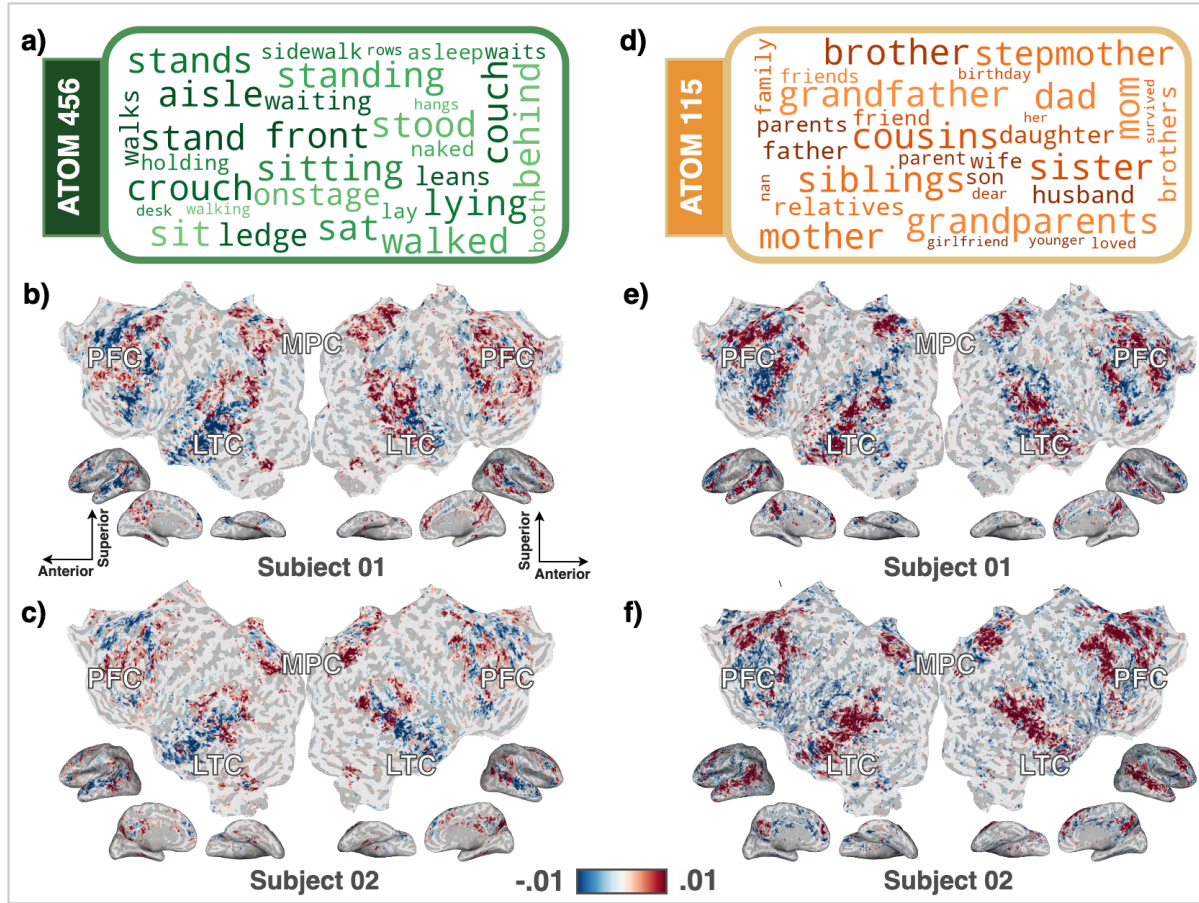


Figure 2.5: The Sparse Concept Encoding Model yields consistent, interpretable cortical maps for individual semantic concepts. (a, d) Two example concept atoms (indices 456 and 115) from the top 20 most strongly activated atoms, interpreted as representing the concepts of “body position” and “family,” respectively. Each word cloud shows the top-activating stimulus words for the corresponding atom. **(b–c, e–f)** Voxelwise regression weights projected onto the selected concept-atom axis for Subject 01 and Subject 02. Red indicates positive selectivity and blue negative. Opacity is scaled by cross-validated prediction accuracy, and only voxels with significant weights (jackknife-estimated $p < 0.05$) are shown. Both atoms exhibit distributed, bilaterally symmetric selectivity patterns that are consistent across individuals. These results show that the Sparse Concept Encoding Model successfully disentangles individual concept representations and produces coherent cortical maps.

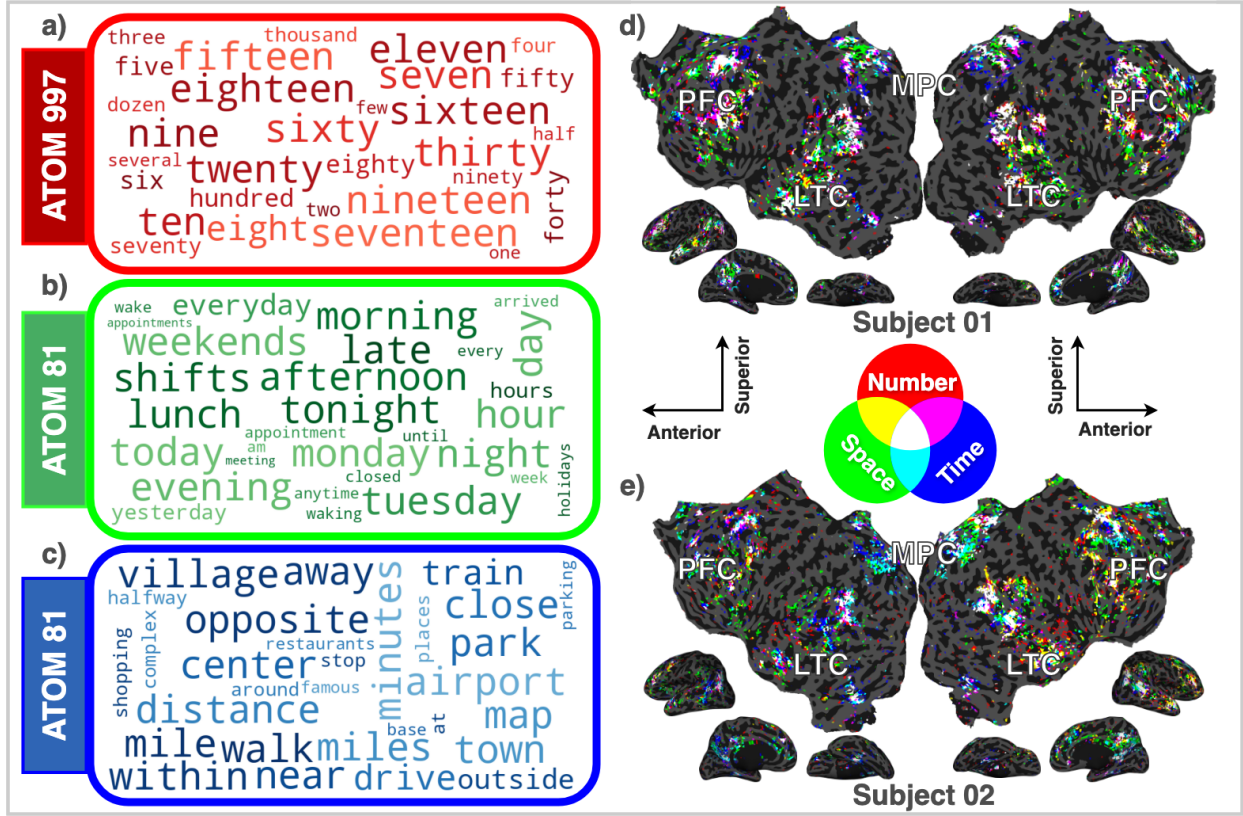


Figure 2.6: Sparse Concept Encoding Model reveals overlapping cortical representations of time, space, and number. (a–c) Word clouds showing the top 30 stimulus words that most strongly activate three selected concept atoms, corresponding to the abstract domains of number (atom 505), time (atom 81), and space (atom 997). (d–e) Composite cortical maps for Subject 01 and Subject 02 showing voxelwise selectivity to each concept atom. Each map displays the top 5,000 voxels with significant positive regression weights (jackknife-estimated $p < 0.05$) per atom. Red, green, and blue denote selectivity to number, time, and space, respectively; overlapping responses are shown in secondary colors, and white indicates joint selectivity to all three. Some inter-individual variation is present, as expected in within-subject functional maps derived from voxelwise models using naturalistic stimuli. These maps reveal localized, bilateral three-way overlap among time, space, and number representations, most notably in the intraparietal sulcus (IPS) and inferior frontal gyrus (IFG). This pattern is consistent with the hypothesis that these domains rely on a shared magnitude system in the human brain.

quantify the extent of overlap, we calculated intersection-over-union (IoU) scores for each concept pair and their three-way intersection (Skagerlund et al., 2016; Cona et al., 2024; Simsek et al., 2025). Averaged across seven participants, IoU scores were 0.34 ± 0.05 for time–space, 0.32 ± 0.07 for time–number, 0.32 ± 0.07 for space–number, and 0.18 ± 0.05 for voxels jointly selective for all three. These findings support the shared magnitude system hypothesis, offering fMRI evidence from naturalistic stimuli enabled by a novel and interpretable modeling framework.

2.6.6 Case Study: Comparing Cortical Representations of Different Concepts

Understanding how the brain encodes relationships between abstract concepts is a central goal in cognitive neuroscience. For example, are logical composition concepts such as "combine", "overlap", and "exclude" represented similarly to body position concepts like "lie," "sit," and "slouch"? To address this, we compared cortical tuning maps across the top 20 concept atoms. For each atom, we generated cortical maps as described in Section 2.6.4 and computed pairwise cosine similarities between all maps, yielding a 20×20 concept-level similarity matrix. We applied hierarchical clustering to the similarity matrix from Subject 01 and visualized the reordered matrix alongside its dendrogram. Using the same atom ordering, we then plotted the similarity matrix for Subject 02. The clustering patterns were qualitatively consistent across subjects and quantitatively stable across all seven subjects: the mean off-diagonal Pearson correlation between similarity matrices across subjects was 0.70 ± 0.08 . To ensure this structure was not inherited from the input embeddings, we computed pairwise Pearson correlations between sparse feature vectors across the 190 concept atom pairs. These correlations were uniformly low ($|r| = 0.00 \pm 0.10$; 85% < 0.10), confirming that the Sparse Concept Encoding Model effectively disentangles semantic directions. In contrast, the same concept directions in the original dense GloVe space showed positive moderate correlations (0.16 ± 0.14 ; 87% > 0), suggesting dense embeddings obscure fine-grained distinctions between semantic directions. These findings suggest that the brain organizes abstract concepts according to a structured similarity geometry, which becomes observable once the feature representations are disentangled.

2.7 Discussions

We introduced the Sparse Concept Encoding Model, a framework that addresses the interpretability challenges arising from superposition in dense feature spaces. By replacing dense superpositions with sparse, overcomplete dictionaries of concept atoms, we reveal each voxel’s tuning along disentangled concept axes. This framework enables new forms of analysis, making it possible to identify shared and distinct cortical representations for related concepts, and to compare representational geometry across concepts.

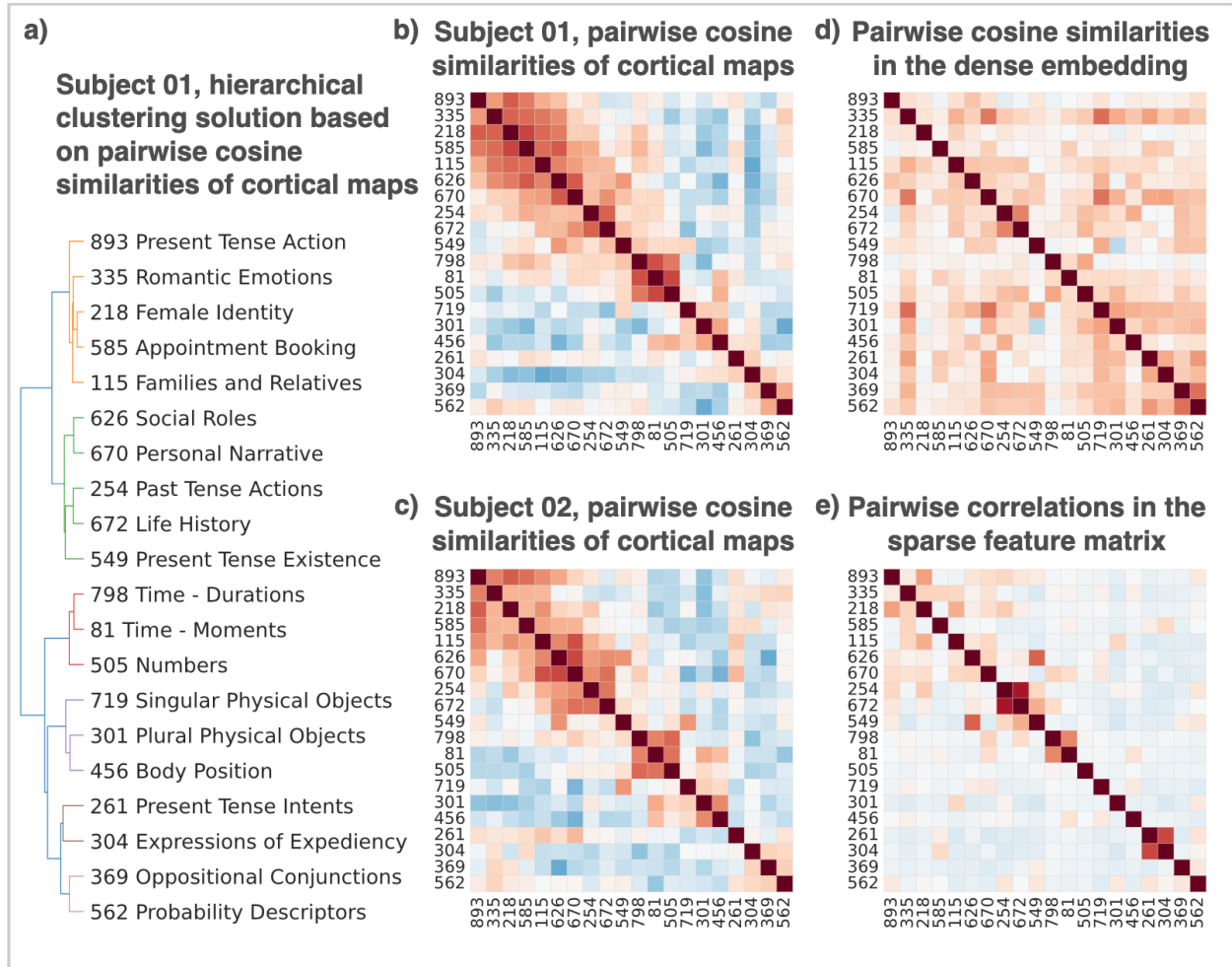


Figure 2.7: Sparse Concept Encoding Model reveals consistent structure in concept-level cortical representations across subjects. (a) Dendrogram from hierarchical clustering of Subject 01 based on pairwise cosine similarities between cortical tuning maps for the 20 most strongly represented concept atoms. Concept labels were heuristically assigned by summarizing the top-activating stimulus words for each atom. (b, c) Concept-level cortical representation similarity matrices for Subject 01 and Subject 02. Each matrix shows pairwise cosine similarity between cortical maps for the 20 concept atoms, reordered using the clustering from Subject 01. The structure is qualitatively consistent across subjects, indicating shared cortical representational structure. (d) Cosine similarity matrix between the same 20 atoms in the original dense GloVe embedding space, showing structured correlations that would confound interpretation under the Dense Encoding Model. (e) Pearson correlation matrix of the sparse feature vectors used to train the Sparse Concept Encoding Model. Off-diagonal values are uniformly low, confirming that the sparse transformation yields disentangled semantic dimensions.

The Sparse Concept Encoding Model plays a role analogous to factorial design in classical cognitive experiments. In naturalistic settings, where orthogonal control is impossible, sparse coding can decorrelate features of interests, making it feasible to isolate effects in uncontrolled conditions.

Critically, the method does not assume that brain responses themselves are sparse. Rather, the L2 penalty is applied when fitting voxel-wise models, encouraging weight vectors to tend to 0. Future work could explore the implications of imposing sparsity directly on voxel weights using L1 regularization.

The number of dictionary atoms used was fixed to 1,000 in this work. While this provides a manageable tradeoff between resolution and computation, increasing this number may allow finer semantic distinctions. Systematically exploring the effect of dictionary size on voxel tuning and interpretability remains an open question.

There have been suggestions that as the number of features in the SAE increases, general features (e.g., "math") tend to "split" into more fine-grained, specialized features (e.g., "algebra", "geometry"). Another interesting idea for future work is to explore continuous "concept manifolds" that blend different concept atoms. It would be interesting to make comparisons between the geometry of the "concept manifolds" and that of the brain's representational space.

This study focused on fMRI data and static word embeddings, but the framework is broadly applicable to domains requiring interpretable mappings from entangled features. One current limitation is that the model does not incorporate linguistic context, as it relies on non-contextual word embeddings. Extending it to contextualized representations from large language models is a promising direction for future work (Jain and Huth, 2018).

Although concept atoms improve interpretability, they may inherit structural biases from the original embeddings, potentially amplifying harmful stereotypes. Moreover, while this method could potentially enhance brain-computer interface technologies by enabling more accurate semantic decoding, it raises ethical concerns around mental privacy and the potential misuse of neural decoding tools (Goering et al., 2021).

2.8 Supplementary Materials

For the main analysis in Section 2.6.5, we selected one representative concept atom from each semantic category—factor 81 (time), factor 505 (number), and factor 997 (space). While our primary results focus on these exemplars, preliminary inspection of other atoms within the same semantic domains suggests they evoke qualitatively similar cortical activation patterns. To assess the similarity of these representations more systematically, we computed cosine similarities between cortical weight maps for all pairwise combinations within three semantic groups: Time (81, 798, 880), Number (394, 505), and Space (353, 997). Number (394–505) and Space (353–997) pairs exhibited consistently high similarity (0.502 ± 0.063 and 0.495 ± 0.088 , respectively), whereas Time pairs showed moderate

consistency (e.g., 81–798: 0.431 ± 0.138 ; 798–880: 0.457 ± 0.098). These within-group similarities were substantially higher than the background distribution of across-group factor pairs, which averaged just 0.048 ± 0.013 . These findings provide quantitative support for the functional clustering of concept atoms within semantic domains and highlight the potential for future research to explore the structure and organization of these representational spaces more deeply.

Table 2.2: Concept atoms with strongest activation for semantic categories of time, number, and space.

Factor Index	Category	Top Associated Words
81	Time	morning, afternoon, evening, weekends, monday, tuesday, lunch, night, hour, day, tonight, today, late, shifts, everyday, yesterday, hours, appointment, am, closed
798	Time	months, hours, days, weeks, ago, years, minutes, hour, month, year, week, day, sec, seconds, spent, intervening, semester, times, last, weekends
880	Time	seventies, sixties, twenties, late, earliest, grunge, subsequent, years, tenure, fame, nineteen, persisted, prior, later, existed, acclaim
394	Number	double, single, two, pair, simultaneous, four, three
505	Number	eleven, sixteen, nineteen, fifteen, seventeen, twenty, ten, eighteen, thirty, sixty, seven, nine, eight, eighty, six, fifty, five, forty, hundred, seventy
602	Number	digs, career, nine, tied, seven, sixth, eight, points, fifth, record, freshman, impressive, five, six, mark, finishes, shy, leads, four, straight
353	Space	threshold, average, calculated, absolute, scale, higher, slope, humidity, altitude, amount, low, normal, mean, distance, lower, level, speed, above, cumulative, readings
997	Space	near, close, distance, airport, center, miles, opposite, minutes, village, mile, town, park, map, away, within, walk, train, drive, outside, halfway

2.8.1 Heuristic Concept-Level Cortical Maps from the Dense Encoding Model

This section presents heuristic concept-specific cortical maps generated from the Dense Encoding Model, used for qualitative comparison against the Sparse Concept Encoding Model, as discussed in Sections 2.6.5 and 2.6.6. Because the Dense Encoding

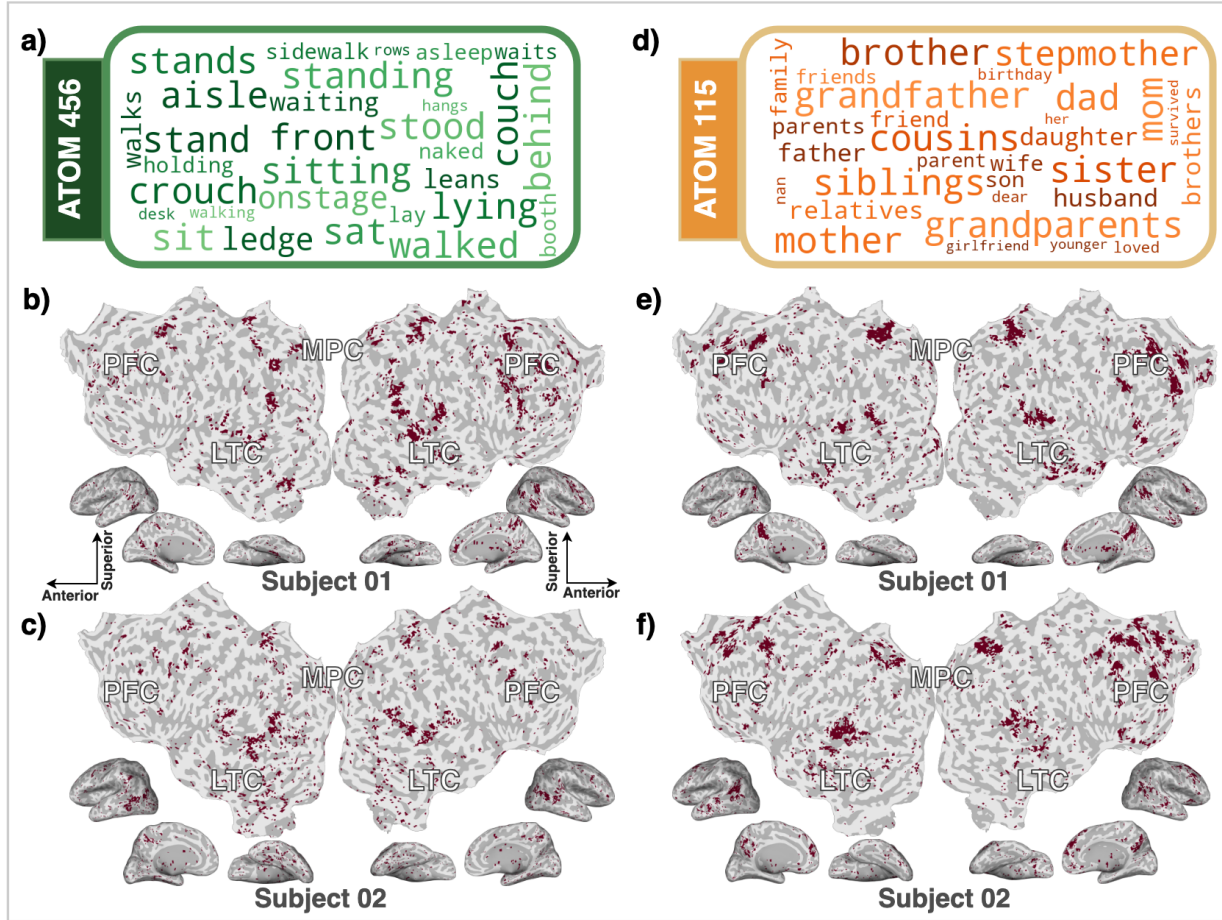


Figure 2.8: Heuristic concept-level cortical maps from the Dense Encoding Model for concept atoms 456 (a–c) and 115 (d–f), shown for Subject 01 and Subject 02. Panels (a, d) display word clouds for the top words associated with each atom. Panels (b–c, e–f) show voxelwise cortical maps, thresholded using the heuristic method described in section 2.8.1. Compared to the corresponding maps in Section 2.6.4 (Fig. 2.5), these maps are less spatially coherent and more variable across subjects, highlighting the interpretability gains achieved by modeling directly in sparse concept space.

Model produces voxel weight vectors in a continuous, entangled semantic space, it does not assign explicit concept axes. To approximate the cortical representation of a specific concept atom, we applied the following heuristic procedure:

Select target concept atom. We began with a target concept atom and retrieved its top 30 activating words using the learned sparse factor embeddings.

Project dense embeddings onto voxel weights. We loaded voxel weights from the Dense Encoding Model. The mean dense embedding of the selected concept atom’s top words was computed and projected onto each voxel’s weight vector, yielding an estimate of that voxel’s tuning strength to the concept.

Select significant voxels. We retained voxels that satisfied both of the following: (a) positive mean cross-validated prediction scores, and (b) a positive projection for the selected concept. This step produced a subset of candidate voxels potentially selective for the concept.

Label voxels by semantic similarity. For each candidate voxel, we computed cosine similarity between its weight vector and all 1,000 sparse concept atoms. We then identified the top 10 most similar atoms for each voxel.

Mask concept-relevant voxels. Voxels were marked as selective for the concept if the original target atom appeared among their top-10 matches.

This procedure was applied independently for each concept examined in Sections 2.6.5 and 2.6.6. While these heuristic maps provide a rough approximation of concept tuning in the dense model, they are less spatially coherent, less symmetric across hemispheres, and more variable across individuals than maps produced by the Sparse Concept Encoding Model. These differences, visualized in Figures 2.8 and 2.9, underscore the interpretability gains afforded by modeling directly in sparse concept space.

2.8.2 Statistical Significance Maps from Permutation Testing

This section presents binary voxelwise significance maps obtained using non-parametric permutation testing to assess prediction accuracy on the test set. To construct a voxelwise null distribution, we performed non-parametric permutation testing. Specifically, the joint predictions of the banded ridge regression model, defined as the sum of predictions from the lexical-semantic and low-level acoustic feature spaces, were randomly permuted over time within each voxel. To account for temporal autocorrelation in the BOLD signal, we used a block permutation strategy following (Deniz et al., 2019), permuting non-overlapping blocks of 10 consecutive time points (equivalent to 20 seconds

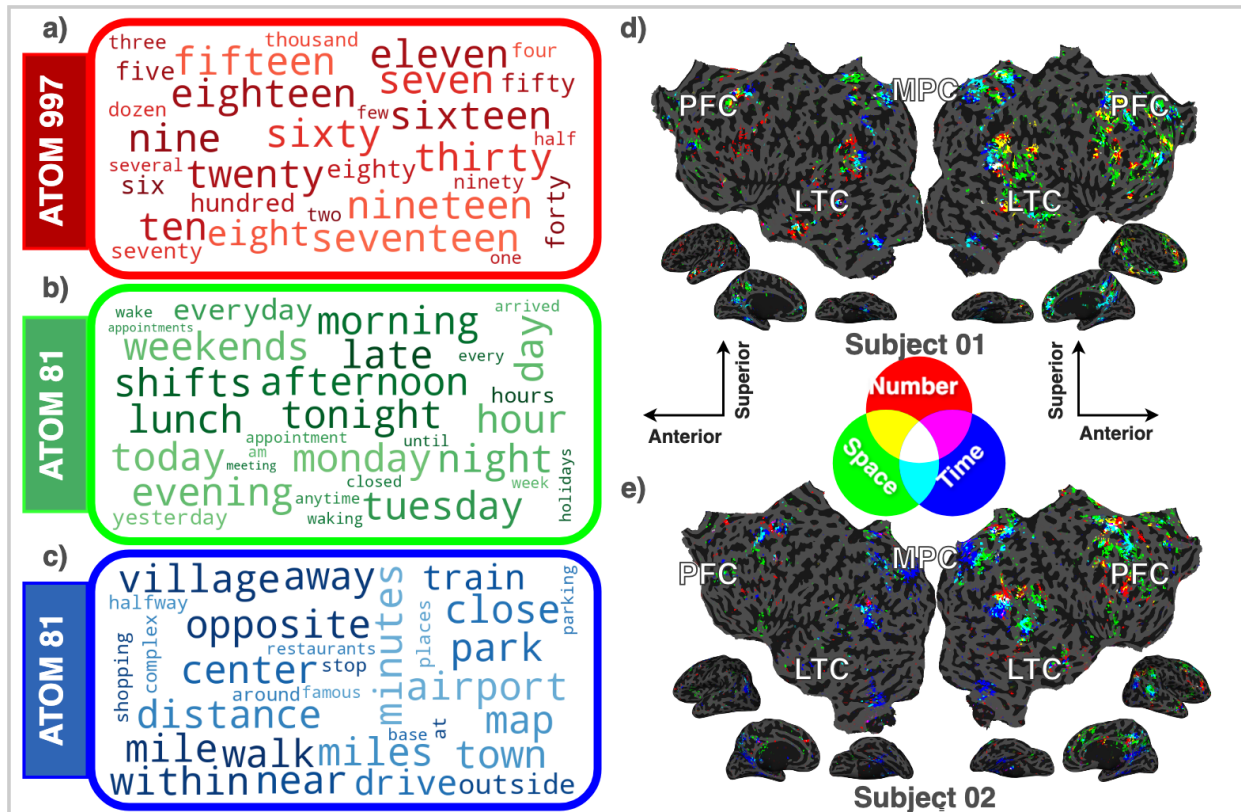


Figure 2.9: Heuristic composite cortical maps from the Dense Encoding Model for three concept atoms representing number (red), time (green), and space (blue). Word clouds (a–c) show the top stimulus words most strongly associated with each concept. Panels (d–e) show voxelwise cortical maps for Subject 01 and Subject 02. Colors indicate voxels selectively tuned to each concept (red, green, blue) or their overlaps (white = joint selectivity). This figure parallels the analysis in Figure 2.6 of the main paper (Section 2.6.5), but reveals substantially less spatial overlap between concepts, highlighting the limitations of the heuristic method.

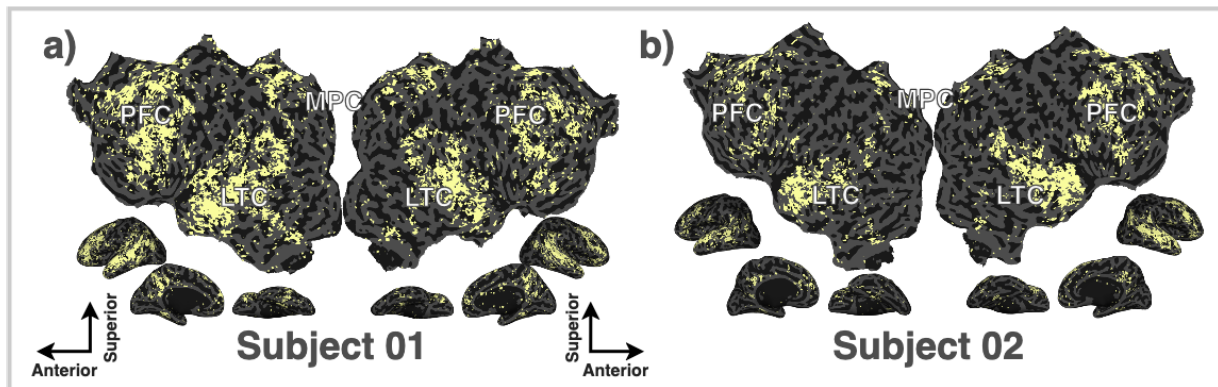


Figure 2.10: Binary statistical significance maps of prediction accuracy for the Sparse Concept Encoding Model in Subject 01 and Subject 02, highlighting significant prediction performance in temporal, parietal, and prefrontal regions.

of data). For each permutation, prediction accuracy was recomputed, and this process was repeated 9,999 times to generate a null distribution of accuracy values for each voxel. An empirical p -value was then computed per voxel by counting the number of permutations in which the permuted prediction accuracy exceeded the observed accuracy from the original (unpermuted) test data. To avoid zero-valued p -values, we added one to both the numerator and the denominator, following the method of (Phipson and Smyth, 2010). To correct for multiple comparisons across voxels, the empirical p -values were adjusted using the Benjamini–Hochberg false discovery rate (FDR) procedure (Benjamini and Hochberg, 1995). FDR correction was applied using the `multiplerests` function from the `statsmodels` Python package with `method='fdr_bh'`. Finally, a binary mask of statistically significant voxels was generated by thresholding the FDR-corrected p -value map at $p < 0.05$.

2.8.3 Evaluation of Embedding Preprocessing Methods

While most learned concept atoms captured coherent semantic themes, we observed that a subset activated broadly across many high-frequency words used in disparate contexts. We hypothesize that this ambiguity stems from a known property of word embeddings: high-frequency words tend to have smaller norms and cluster near the origin in Euclidean space, reducing their separability. To address this, we developed a custom preprocessing method called **Vector Norm Reparameterization (VNR)**. VNR improves the geometric structure of the embedding space by nonlinearly rescaling each word vector's norm while preserving its direction (Fig. 2.11). For a word vector $\mathbf{v}$, we define the scale

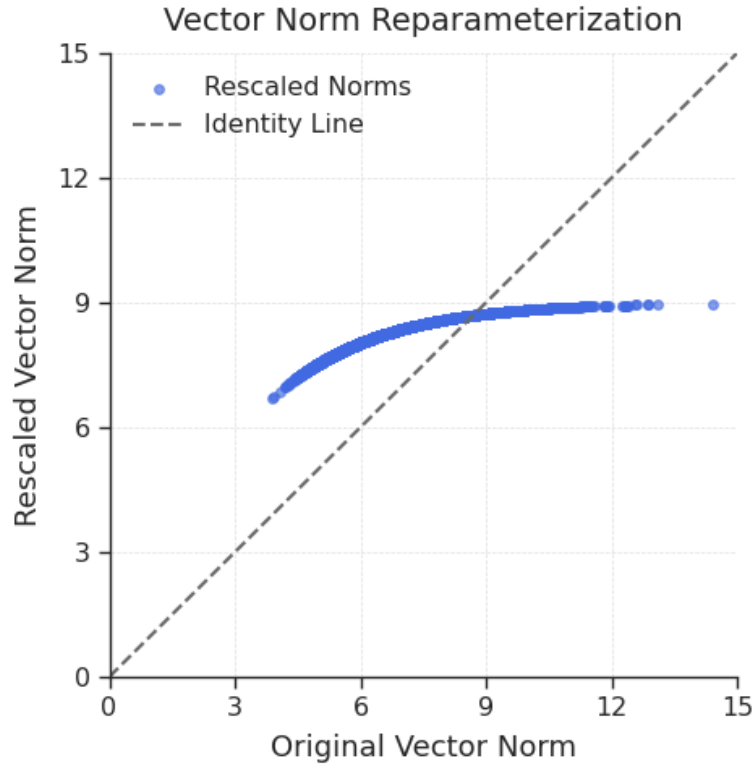


Figure 2.11: Effect of VNR on embedding norms. Each point represents a word vector before and after applying VNR. The dashed line indicates the identity (no change). VNR is a preprocessing method that enhances the interpretability of learned concept atoms by expanding the relative spread of vectors near the origin while compressing those among high-norm vectors.

factor as:

$$s = 9 \cdot \tanh \left(\frac{2\|\mathbf{v}\|}{9} + \frac{1}{10} \right),$$

and the reparameterized vector becomes:

$$\tilde{\mathbf{v}} = \frac{\mathbf{v}}{\|\mathbf{v}\|} \cdot s.$$

This transformation compresses large norms more than small ones, effectively expanding semantically dense regions and reducing the influence of rare words.

To evaluate the effectiveness of VNR, we compared it against three alternative preprocessing strategies applied prior to dictionary learning:

1. Unmodified GloVe vectors
2. L2 normalization to the unit hypersphere
3. ZCA whitening

Each method was evaluated using the following metrics:

Reconstruction Error. We computed the relative Frobenius norm $\|X - DZ\|_F / \|X\|_F$, where $X \in \mathbb{R}^{300 \times N}$ contains the preprocessed word vectors, $D \in \mathbb{R}^{300 \times 1000}$ is the learned dictionary, and $Z \in \mathbb{R}^{1000 \times N}$ is the corresponding sparse representation. This metric measures how well the sparse coding model reconstructs the transformed embedding space.

Sparsity. For each word, we calculated the fraction of zero entries in its 1,000-dimensional sparse code, then averaged this value across the stimuli vocabulary. Higher values indicate sparser representations.

WordNet Purity. To assess semantic coherence, we identified the top 50 highest-activation words for each concept atom and computed the mean pairwise WordNet similarity among them, using both path similarity and Wu–Palmer similarity (Miller, 1995; Pedersen et al., 2004; Wu and Palmer, 1994). Semantically meaningful atoms are expected to group together words from similar lexical fields, as defined in WordNet’s ontology.

Results. As shown in Table 2.3, VNR achieved the lowest reconstruction error while maintaining high sparsity and the strongest semantic coherence across both WordNet-based metrics. Notably, L2 normalization resulted in extremely poor reconstruction performance, likely due to its constraint that all word vectors and dictionary atoms lie on the unit hypersphere, limiting the model’s ability to encode useful norm variation. Based on these results, we adopted VNR as the default preprocessing method in all subsequent experiments.

Table 2.3: Comparison of preprocessing strategies for word embeddings prior to sparse coding.

Embedding Version	Reconstruction Error	Sparsity	WN Purity (Path)	WN Purity (WuP)
Raw GloVe	0.4922	0.9723	0.1158	0.3571
VNR	0.4083	0.9677	0.1171	0.3579
L2-normalized	7.2403	0.9640	0.1166	0.3563
ZCA-whitened	0.6714	0.9610	0.1113	0.3412

CHAPTER 3

Deep Brain Stimulation Amplifies Movement-Related Neural Modulation in Parkinson's Disease: Evidence from Multimodal Home Recordings

3.1 Abstract

Parkinson's disease (PD) is a dynamical disease: motor symptoms arise not from static lesions but from aberrant neural activity across basal ganglia and sensorimotor cortex. Deep brain stimulation (DBS) is effective but remains open-loop, with parameters tuned heuristically. Here we use a home-deployed multimodal platform to quantify how DBS amplitude modulates movement-related neural activity. Two participants implanted with bidirectional neurostimulators (Medtronic Summit RC+S) completed six recording days at home, performing a UPDRS-derived motor battery at three stimulation amplitudes (under-, preferred-, and over-stimulation). We recorded local field potentials from subthalamic nucleus and electrocorticography from precentral and postcentral gyri, alongside bilateral wrist accelerometry and video data for pose estimation. Encoding models predicted neural spectral power from behavioral features: mean acceleration, full acceleration spectrogram, task labels, pose kinematics, and a combined set. Across participants, hemispheres, and recording sites, behavioral features robustly predicted neural activity; temporally rich representations—full acceleration spectrogram and task labels—substantially outperformed coarse summaries, indicating that fine-grained movement structure is necessary for capturing behaviorally driven neural variance. In the participant with stronger model performance, higher stimulation amplitudes enhanced prediction throughout the sensorimotor network by amplifying a movement-locked spec-

tral component characterized by beta suppression and gamma enhancement. This pattern aligns with “information lesion” accounts of DBS in which stimulation restores movement-related signaling through pathologically synchronized circuits, and provides an empirical basis for adaptive neuromodulation informed by behavioral state.

3.2 Introduction

Real-time feedback control is routine in many complex engineered systems, from aircraft autopilots to biomedical devices such as rate-adaptive cardiac pacemakers and closed-loop insulin pumps. Comparable closed-loop control has not yet been achieved for neurological disease. Deep brain stimulation (DBS) for Parkinson’s disease (PD) is a prime example: more than two decades after its introduction, DBS remains an open-loop therapy. Standard clinical practice relies on fixed, high-frequency monopolar pulse trains delivered to basal ganglia targets via an implanted pulse generator, with stimulation parameters tuned heuristically based on clinician experience rather than derived from explicit models of neural dynamics or behavioral state (Schiff, 2010; Gorzelic et al., 2013; Koeglsperger et al., 2019).

A central factor limiting progress toward principled, adaptive DBS is the dynamical complexity of the underlying disease. PD has been characterized as a “dynamical disease” because motor symptoms arise from aberrant activity patterns within distributed neuronal networks rather than from static lesions (Beuter and Vasilakos, 1995). The relevant system includes nonlinear interactions among basal ganglia, motor cortex, cerebellum, and brain stem circuits, all driven by continuously changing sensory, motor, and environmental context (Brown, 2003; McGregor and Nelson, 2019; Grill et al., 2004; Shenoy et al., 2013; Caligiore et al., 2016). Therapeutic optimization therefore demands both a systems-level understanding of disease mechanisms and computational models capable of capturing time-varying, nonlinear brain–behavior dynamics.

Only recently has the requisite technology matured. Bidirectional neurostimulators such as the investigational Medtronic Summit RC+S can now deliver therapeutic stimulation while simultaneously recording local field potentials chronically in freely moving patients (Stanslaski et al., 2018; Starr, 2018). In parallel, wearable inertial sensors and camera-based pose estimation enable dense, time-resolved behavioral measurement in naturalistic settings (Deng et al., 2024; Strandquist et al., 2023). Together, these advances establish the technical foundation for integrating multimodal, high-volume data streams into interpretable models of brain–behavior relationships.

The present study leverages a multimodal data-collection platform developed by the Gallant Lab in collaboration with UCSF and the University of Washington that combines an implanted Summit RC+S neurostimulator with wrist-worn accelerometers and home-deployed video cameras (Strandquist et al., 2023; Dixon et al., 2025). Two participants completed six recording days at home, performing a fixed battery of motor tasks at three

monopolar stimulation amplitudes spanning under-, preferred-, and over-stimulation states. Within this framework, encoding models were used to predict neural spectral activity from behavioral features. Prediction performance was compared systematically across feature sets (mean acceleration, full acceleration spectrogram, task labels, and pose kinematics) and stimulation conditions.

This design introduces several methodological advances. First, rather than contrasting gross movement against rest at a single stimulation setting, we sample a range of semi-naturalistic motor states across multiple stimulation amplitudes. This design addresses two questions: which aspects of movement are encoded in which frequency bands, and how stimulation modulates the fidelity of this encoding. Second, we collect high-volume, within-individual datasets and evaluate models using modern cross-validated pipelines on held-out data from the same individual. This reduces the risk of inflated type I error from group-level averaging and preserves meaningful individual differences. Third, data collection occurs primarily in the home environment, with remote monitoring and secure data transfer after an initial in-person setup. This demonstrates the feasibility of studying complex stimulation–movement interactions under everyday conditions.

Three main findings emerge. First, behavioral features robustly predict neural spectral power across a broad frequency range in STN, precentral gyrus, and postcentral gyrus. Features with richer temporal and categorical structure—particularly full acceleration spectrogram and task labels—explain substantially more variance than coarse summaries such as mean acceleration, indicating that fine-grained movement structure is necessary for capturing behaviorally driven neural variance. Second, in the participant with stronger prediction performance, preferred- and over-stimulation markedly enhance prediction performance throughout STN and sensorimotor cortex. This effect is evident in both hemispheres, each driven by an independent neurostimulator and modeled separately. Third, this enhancement reflects amplification of a movement-locked spectral component spanning beta, gamma, and alpha/delta bands, including a prominent peak at stimulation-entrained gamma. This stimulation-dependent amplification of movement-related neural modulation accords with “information lesion” accounts of how DBS reshapes pathological dynamics and, to our knowledge, has not previously been demonstrated in chronic human recordings.

3.3 Background

3.3.1 Parkinson’s Disease and Deep Brain Stimulation

PD is the second most prevalent neurodegenerative disorder worldwide, affecting more than 10 million people (Ou et al., 2021). It is characterized by the progressive loss of dopaminergic neurons in the substantia nigra pars compacta, leading to substantial dopamine depletion in the striatum. This depletion disrupts the balance between the

direct and indirect motor pathways of the basal ganglia (Albin et al., 1989; McGregor and Nelson, 2019) and produces the core motor symptoms of PD: akinesia, rigidity, and tremor. As the disease progresses, patients often develop gait and postural instability, cognitive impairment, and affective disturbances (Poewe et al., 2017).

The primary treatment for PD is pharmacological. Most patients initially receive dopamine replacement therapy, such as carbidopa–levodopa, or dopamine receptor agonists, such as pramipexole. These medications are often highly effective in the early stages of the disease. Over time, however, many patients develop symptoms that are refractory to pharmacological management. Common complications include levodopa-induced dyskinesias and motor “on–off” fluctuations, in which periods of good motor control alternate with periods of severe motor impairment (Thanvi et al., 2007; Connolly and Lang, 2014). When these complications can no longer be managed by medication adjustments alone, patients are often evaluated for neurosurgical intervention.

DBS was first demonstrated as an effective treatment for PD by Alim-Louis Benabid and colleagues in the late 1980s. Their work demonstrated that high-frequency stimulation could reproduce many of the benefits of prior treatments like ablative surgery while remaining reversible and adjustable (Williams, 2010). DBS has since become a standard therapy for advanced PD (Lozano et al., 2019). In current practice, DBS delivers high-frequency (typically 100–200 Hz) electrical pulses through implanted electrodes targeting basal ganglia nuclei such as the subthalamic nucleus or the internal segment of the globus pallidus. Clinical trials and long-term follow-up studies have confirmed that DBS can improve tremor, rigidity, and bradykinesia, reduce motor fluctuations and dyskinesias, and enhance quality of life.

3.3.2 Hypothetical Mechanisms of DBS

Despite widespread clinical use, the mechanisms by which DBS produces therapeutic benefit remain incompletely understood. Several hypotheses have been proposed, differing in emphasis but broadly complementary (Lozano et al., 2019; Ashkan et al., 2017).

Classic Rate Model The earliest explanatory framework for DBS was the inhibition hypothesis, grounded in the classic rate model of basal ganglia function. In PD, degeneration of dopaminergic neurons in the substantia nigra pars compacta leads to elevated firing rates in the subthalamic nucleus (STN) and internal segment of the globus pallidus (GPi). According to the rate model, this hyperactivity strengthens inhibition of thalamic relay nuclei, reduces excitatory drive to motor cortex, and thereby contributes to akinesia and rigidity. Within this framework, high-frequency DBS was thought to act as a “functional lesion”—suppressing pathological hyperactivity in the target nucleus, relieving excessive downstream inhibition, and restoring more normal thalamocortical output. This hypothesis aligned with observations that DBS produced effects similar to anatomical

lesioning (Albin et al., 1989; DeLong, 1990; Bergman et al., 1990). Subsequent electrophysiological studies, however, indicated that this account is incomplete. High-frequency stimulation of STN or GPi does not consistently silence local neurons; somatic firing can instead increase during stimulation and become entrained to the pulse train (Hashimoto et al., 2003; Ashkan et al., 2017). These findings contradict the simple inhibition hypothesis and support an alternative view in which DBS reshapes firing patterns and information flow across the network.

Disruption and Replacement of Pathological Oscillations A second major hypothesis emphasizes abnormal oscillations and synchrony in PD. Local field potential and single-unit recordings in both humans and animal models display elevated beta-band power ($\sim 13\text{--}35$ Hz), increased bursting, and abnormal synchronization throughout cortico-basal ganglia circuits (Yu et al., 2021; Stein and Bar-Gad, 2013; Little et al., 2012; Pogosyan et al., 2010). In healthy subjects, beta oscillations are associated with maintaining the current motor state. When beta synchrony becomes excessive and persistent, it is thought to hinder movement initiation and contribute to bradykinesia and rigidity (Leventhal et al., 2012; Brittain and Brown, 2014; Khanna and Carmena, 2017; Brown, 2003). Within this framework, DBS acts primarily by disrupting or replacing pathological beta oscillations and coherent firing patterns. High-frequency stimulation reduces beta power and desynchronizes abnormally coupled nuclei, promoting a more flexible, less synchronized regime that better supports voluntary movement. This network-level regularization of activity is sometimes described as an “informational lesion.” (Grill et al., 2004; Rubin and Terman, 2004; Spooner et al., 2024)

Multiscale Neuromodulation and Plasticity More recent perspectives characterize DBS as multiscale neuromodulation, engaging distinct mechanisms across spatial and temporal scales (Neumann et al., 2023; Ashkan et al., 2017). A key observation motivating this view is that different symptom categories improve on distinct timelines. Tremor and rigidity can resolve within seconds to minutes of stimulation onset, consistent with rapid changes in spike timing and local field potentials. Bradykinesia often improves over tens of minutes to hours, suggesting slower network reconfiguration or neuromodulatory processes. When DBS is applied to psychiatric indications such as obsessive-compulsive disorder or depression, clinically meaningful changes typically emerge only after weeks to months, implicating longer-term plasticity. Under this framework, DBS does not operate through a single mechanism but rather engages multiple physiological processes across different timescales and distributed brain circuits (Ashkan et al., 2017).

3.3.3 Limitations of Current DBS and Directions for Improvement

DBS is effective for PD but does not provide complete relief and can produce substantial side effects. Many patients continue to experience gait impairment, speech difficulties, mood changes, or dyskinesias even under “optimized” settings. Long-term stimulation also entails ongoing device maintenance and regular recharging. These limitations stem from two features of current practice: how stimulation is programmed and how it is delivered over time. Accordingly, two complementary improvement strategies have emerged: refining how stimulation is delivered and refining when it is delivered. Both aim to increase efficacy, reduce side effects, and move DBS toward more precise, individualized neuromodulation.

Refining How Stimulation Is Delivered The first strategy targets the structure and spatial targeting of stimulation within a high-dimensional parameter space. Modern DBS systems permit adjustment of contact configuration, polarity, amplitude, frequency, and pulse width, yet routine clinical practice explores only a small subset of this space. Standard programming relies on regular high-frequency pulse trains near 130 Hz, short pulse widths (typically 60–90 μ s), and amplitudes of 1–4 V (or mA), delivered through one or a few contacts surrounding the target nucleus. These conventional settings cover only a narrow slice of the possible stimulation space.

A major reason for this underutilization is the manual, heuristic nature of parameter selection. Parameters are tuned through iterative clinical testing without widely adopted data-driven protocols. Optimization thus remains a time-consuming trial-and-error process that depends heavily on clinician experience and requires repeated in-person visits—particularly burdensome for patients far from specialized centers. These constraints underscore the need for systematic, data-driven exploration of the parameter space guided by objective neural and behavioral measures (Koeglsperger et al., 2019).

Refining When Stimulation Is Delivered The second strategy entails closing the loop between neural or behavioral state and stimulation output. Conventional DBS (cDBS) delivers stimulation continuously with fixed parameters, regardless of fluctuations in neural activity, medication state, or symptoms. This “always-on” approach often mismatches stimulation to momentary clinical needs: periods of relative wellness receive as much stimulation as periods of severe impairment. Continuous delivery can exacerbate stimulation-induced side effects and accelerate battery depletion.

Adaptive DBS (aDBS) addresses these limitations by adjusting stimulation in real time based on feedback signals reflecting symptom severity or network state. Stimulation parameters are modulated according to neural biomarkers (for example, STN beta power, stimulation-entrained gamma) and/or behavioral measures (for example, accelerometer-derived kinematics). Rather than delivering constant output, aDBS increases stimulation

when biomarkers indicate worsening symptoms and decreases or suspends it when the patient's state is relatively stable. Multiple studies have demonstrated that aDBS can match or exceed the motor benefits of cDBS while reducing total stimulation time, energy delivered, and side effects (Little et al., 2016; Oehrns et al., 2024).

3.3.4 Neural Biomarkers for Parkinson's Disease

A central goal in PD research is to identify biomarkers that index symptom state in real time. Such markers are essential for data-driven selection of stimulation parameters and for adaptive DBS. This requires signals, neural or behavioral, that reliably track pathological circuit dynamics. Most work has focused on neural signals recorded directly from implanted DBS electrodes, which are continuously available and provide a direct window onto activity at the stimulation target (Little and Brown, 2012, 2014).

The most established biomarker is beta-band oscillatory activity ($\sim 13\text{--}30$ Hz) in the basal ganglia (Little and Brown, 2014). Beta power in STN rises during bradykinesia and rigidity and falls following dopaminergic medication or DBS onset (Khanna and Carmena, 2015; Yu et al., 2021; Stein and Bar-Gad, 2013; Little et al., 2012; Pogosyan et al., 2010). Because these fluctuations parallel clinical state, beta power is a natural candidate for feedback control. Adaptive paradigms that trigger stimulation when beta power exceeds a threshold have demonstrated that motor symptoms can be controlled—and in some cases more effectively—while delivering substantially less total stimulation than conventional continuous DBS (Little et al., 2016).

A second, more recently characterized biomarker is stimulation-entrained gamma. In the absence of stimulation, narrowband gamma oscillations emerge in STN and motor cortex during high-dopamine states. During active DBS, the pulse train can entrain this rhythm to a subharmonic of the stimulation frequency, typically in a 1:2 ratio (e.g., ~ 65 Hz gamma during 130 Hz stimulation) (Sermon et al., 2023; Olaru et al., 2025). Entrained gamma amplitude covaries with medication state and motor function, positioning it as another candidate for closed-loop control (Olaru et al., 2024; Mathiopoulou et al., 2025). Adaptive strategies using entrained gamma as the feedback signal have achieved improved management of motor fluctuations and dyskinesias relative to continuous stimulation (Oehrns et al., 2024).

3.3.5 Bidirectional Implantable Neurostimulators

Systematic, data-driven optimization of DBS requires dense neural recordings paired with objective measures of behavioral state. Much current understanding of PD derives from recordings obtained via externalized cortical or subcortical leads, either during implantation or in the immediate postoperative period. These data are invaluable because they offer high signal-to-noise ratio and excellent spatiotemporal resolution compared with noninvasive methods. However, they are inherently limited: recordings are brief, typically

acquired at rest or during simple movements, and collected in an artificial hospital setting where neural activity is likely altered by recent surgery (Starr, 2018).

These constraints motivated development of fully implanted devices capable of both stimulating and sensing neural activity chronically under naturalistic conditions. Such bidirectional interfaces integrate recording and stimulation within a single implant, enabling long-term monitoring of neural dynamics across everyday contexts and stimulation states. The investigational Summit RC+S system (Medtronic) was designed for this purpose (Stanslaski et al., 2018; Gilron et al., 2021). It incorporates hardware and firmware features that facilitate sensing by reducing stimulation artifacts during active stimulation, including configurable blanking windows around each pulse and an active-recharge mode that shortens the recovery phase. The device streams neural data wirelessly to an external tablet while patients move freely, supporting multi-day recordings during naturalistic behavior (Stanslaski et al., 2018; Ansó et al., 2022). The present study leverages this bidirectional platform to characterize how stimulation amplitude, neural oscillations, and movement interact under real-world conditions.

3.4 Methods

3.4.1 Experimental Design

Two participants with Parkinson's disease were recruited from an adaptive DBS trial at the University of California, San Francisco (UCSF) approved by the Institutional Review Board (IRB) and conducted under an Investigational Device Exemption (IDE); both participants were enrolled after providing written informed consent.

Each participant had a pair of electrode leads in each hemisphere: a quadripolar depth electrode targeting the STN and a quadripolar subdural ECoG strip spanning the arm/hand representations of the precentral and postcentral gyri (primary motor and primary somatosensory cortex). Each lead pair was connected to an investigational, sensing-enabled neurostimulator (Summit RC+S, Medtronic) implanted in the ipsilateral subclavicular space. The RC+S device delivered therapeutic stimulation to the STN, recorded LFPs from the STN and ECoG contacts, and recorded three-axis accelerometry from an inertial sensor in the device case. Participants could revert to their standard clinical DBS settings at any time using a handheld patient programmer, ensuring safety and comfort throughout the study.

The study was designed to collect synchronized neural, wearable, and video data in the home environment. The aim was to sample a range of motor states under systematically varied stimulation amplitudes, in conditions that were controlled yet ecologically valid, and to relate these states to multichannel neural activity. To our knowledge, no previous work has combined chronic RC+S recordings with multi-view video and bilateral wrist-worn inertial measurements in the home while explicitly probing multiple movement states

across multiple stimulation levels. Stimulation frequency was held at each participant's usual clinical settings (130 Hz for Patient 01 and 150 Hz for Patient 02).

Each participant completed six recording days. On each day, the participant performed a fixed battery of motor tasks at three monopolar stimulation amplitudes. These amplitudes were chosen to span distinct clinical states—under-stimulated, clinically preferred, and over-stimulated—and were held constant within each task block. Each recording day comprised three stimulation blocks, with the order of amplitudes permuted in a counter-balanced sequence across days to reduce potential confounds from time-of-day, fatigue, or task adaptation. Daily sessions lasted approximately 25 minutes and were scheduled at similar times of day across days to minimize circadian and medication-related variability.

The task battery was based on standardized items from the Unified Parkinson's Disease Rating Scale (UPDRS) Part III ([Goetz et al., 2008](#)), supplemented with additional functional tasks. Participants received a written instruction sheet specifying the task order and were trained to perform all tasks before home recording began. During each session, participants completed the following self-guided tasks:

- Rest: seated rest with hands supported, used to assess baseline tremor and dyskinesia.
- Finger tapping: repetitive index–thumb opposition.
- Hand opening–closing.
- Hand pronation–supination.
- Sit-to-stand and standing balance.
- Walking within the home office space.
- Foot stomping.
- Typing: copying short blocks of text (approximately 100 words) on their home computer.

The full sequence of tasks was repeated at each of the three stimulation amplitudes on each day, yielding multiple repetitions of each task at different stimulation levels. Full patient instructions are provided in Supplementary Materials [3.7.1](#).

3.4.2 Data Collection Platform

A remote data collection platform was installed in each participant's home office to support multimodal recording and remote monitoring. The platform integrated four primary data streams: intracranial neural activity from the Summit RC+S device, wrist

accelerometry from bilateral smartwatches, video, and, for Patient 01 only, keystroke-based measures of typing performance. After installation, the system was operated and maintained remotely and was used to collect data during both structured tasks and free behavior.

3.4.2.1 RC+S Recordings

All neural recordings were performed during monopolar stimulation. In the STN, LFPs were recorded using a symmetric bipolar “sandwich” configuration, with the recording contacts placed on either side of the stimulating contact to enhance common-mode rejection of stimulation artifacts. Along the ECoG strip, two bipolar channels were recorded: one spanning electrodes over the postcentral gyrus and one spanning electrodes over the precentral gyrus. For all three channels, the RC+S was configured with a sampling rate of 500 Hz, a high-pass filter at 0.85 Hz, and a low-pass filter at 100 Hz. A sense-blanking period of 2.5 ms was applied around each stimulation pulse to further reduce artifacts. The device also streamed actigraphy from its embedded accelerometer at 64 Hz, providing an additional inertial signal synchronized with the neural recordings.

3.4.2.2 Wearable Accelerometry

Arm movement was measured using Apple Watches worn on each wrist. Three-axis accelerometer data were sampled at 50 Hz and recorded continuously during each session. Recordings were automatically uploaded to a HIPAA-compliant third-party server (Rune Labs, Inc.), from which they were later downloaded for analysis.

3.4.2.3 Video Recordings

Video recordings were used to capture whole-body movement and to provide input for downstream pose estimation. For Patient 01, three cameras were installed in the home office at different viewpoints. Cameras were positioned to maximize coverage of the upper body and workspace during seated and standing tasks. Video was recorded at a constant frame rate of 30 frames per second, and a timestamp was stored for every frame. A custom Java application controlled the recording schedule via a small touchscreen interface. This graphical user interface allowed the participant to start and stop recordings and to cancel scheduled recordings without needing a keyboard or mouse. Video files were first stored on the local acquisition machine and then transferred securely. For Participant 02, video data were collected remotely via Zoom during supervised sessions. Zoom recordings were captured at 25 frames per second.

3.4.2.4 Data Synchronization

Accurate time alignment across data streams is essential to relate neural activity to behavior. In our setup, all devices (Summit RC+S, Apple Watches, and cameras) maintained their individual internal clocks, so raw timestamps were not initially synchronized. To address this, we implemented a simple synchronization gesture that participants performed at the start of each recording block.

At the beginning of each session, the participant was instructed to tap the implantable neurostimulator (INS) with one hand while keeping both hands visible in all camera views. Direct contact with the INS produced a sharp transient in the RC+S accelerometer signal, a concurrent high-amplitude event in the Apple Watch accelerometers, and a brief, easily identifiable motion in the video-derived hand trajectories. These synchronized transients served as a common temporal landmark across modalities.

For each recording segment, we manually aligned the neural, wearable, and video data by matching this synchronization event using a publicly available analysis tool `ManualTimeAlignerGUI`¹. This procedure yielded a common reference time for all devices and provided sample-level alignment between neural and behavioral signals.

3.4.2.5 Data Transfer

All devices were connected to a central acquisition computer in the participant's home office. The acquisition computer communicated with institutional servers over a virtual private network (VPN). Data were encrypted in transit and at rest. The platform supported remote monitoring of system status and remote troubleshooting, minimizing the need for in-person visits and enabling repeated data collection across days with low participant burden.

3.4.3 Stimulation Amplitude Titration

Stimulation amplitudes were titrated before data collection to define three clinically meaningful levels for each participant. Titration sessions were conducted on separate days, using bradykinesia, tremor, and dyskinesia as clinical feedback in consultation with the study neurologist. For each participant, the clinically preferred monopolar stimulation amplitude served as the reference. The under-stimulation level was set to 50% of this preferred amplitude, and the over-stimulation level was set to 110% of the preferred amplitude, subject to tolerability constraints. These three amplitudes (under, preferred, over) were then used as the three stimulation conditions in the experimental sessions.

¹<https://github.com/Weill-Neurohub-OPTiMaL/ManualTimeAlignerGUI.git>

3.4.4 RCS Time Series Preprocessing

The Summit RC+S neurostimulator streams neural data in discrete UDP packets encoded as JSON. Each packet carries an internal timestamp and sequence counters. Packets are occasionally lost in transmission, resulting in dropped data that must be handled. The device also provides a 16-bit tick counter (100 μ s resolution) that rolls over every ~ 6.553 s, along with a system timestamp (in seconds) for each packet. A custom preprocessing algorithm, extending the OpenMind Consortium's Analysis-rcs-data toolbox², was used to merge these packets into continuous, uniformly sampled time series (Sellers et al., 2021).

Anomalous packets were first removed, including those with negative or duplicate timestamps or unexpected sample counts. Continuous recording segments were then identified by detecting gaps in the device timing signals. Successive packet timestamps, system ticks, and data sequence numbers were compared against their expected increments, with corrections for tick counter rollover. A breakpoint was marked whenever a timestamp or counter jumped by more than one packet interval, indicating a discontinuity. Gaps were categorized by duration. If a gap was shorter than the tick counter overflow period (~ 6.553 s), the surrounding data were treated as a single segment, and the system tick difference was used to estimate the missing time and preserve continuity. Larger gaps were treated as boundaries between segments, and a new segment was started. For each continuous segment, a corrected time base was established using the device generation timestamps combined with cumulative tick counts. Within each packet, samples were assumed to be evenly spaced at the device sampling rate. This procedure produced a cleaned, chronologically ordered time-series pandas DataFrame indexed by timestamp, composed of gap-free segments of neural data suitable for feature extraction.

3.4.5 Neural Spectral Feature Extraction

For each continuous segment of the 500 Hz neural time series, a moving-window spectrogram was computed to quantify oscillatory power over time. A short-time Fourier transform with 1 s Hanning windows and 50% overlap (`scipy.signal.spectrogram`) yielded one power spectrum every 0.5 s. Spectral power was estimated from 0 Hz to the 250 Hz Nyquist frequency with 1 Hz resolution for each recorded neural channel. To obtain log-power spectra, values were transformed using an epsilon-stabilized logarithm ($\text{power} + 1 \times 10^{-8}$), and extreme values were clipped to lie within three standard deviations of the mean. The resulting neural spectral features form a time series of log-power spectra, where each 0.5 s bin contains power in 251 frequency bands (0–250 Hz) for each recorded channel.

²<https://github.com/openmind-consortium/Analysis-rcs-data>

3.4.6 Watch Accelerometer Spectral Feature Extraction

An analogous procedure was applied to the wrist accelerometer data to capture movement- or symptom-related power. The raw triaxial accelerometer signal, sampled at 50 Hz, was divided into 3s Hanning windows with 2.5s overlap, providing one power spectrum every 0.5s aligned to the neural features. A 3s window was chosen to improve resolution of low-frequency movements. Spectral power was estimated from 0 to the 25 Hz Nyquist frequency with 0.33 Hz resolution for each axis (X, Y, Z). Power values were then converted to log-power using an epsilon-stabilized logarithm ($power + 1 \times 10^{-8}$), and extreme values were clipped to lie within three standard deviations of the mean. The resulting accelerometer spectral features form a time series of log-power spectra, where each 0.5 s bin contains power across 76 frequency bands (0–25 Hz) for each of the three axes.

3.4.7 Task Label Feature Extraction

We reviewed the video for each session and manually marked the beginning and end of each instructed task. These annotations were converted into a time-resolved task label sequence spanning the entire recording. For modeling, task labels were one-hot encoded into a binary matrix with columns corresponding to specific tasks where ones indicate active task performance and zeros indicate otherwise. This procedure produced a set of task indicator time series that were aligned with the neural, accelerometer, and pose features.

3.4.8 Pose Feature Extraction

Synchronized video recordings were processed to extract kinematic features of the upper body. A publicly available pose estimation algorithm (MediaPipe)³ was used to obtain 3D coordinates of key body landmarks, including shoulders, elbows, wrists, fingertips, and hips (Lugaresi et al., 2019). The video sampling rate was 30 Hz for Patient 01 and 25 Hz for Patient 02. Lower-limb landmarks were excluded, as participants were seated for all tasks. From the landmark coordinates, joint-angle time series were computed to quantify upper-limb movements. For each frame, the angle for a given joint was defined by three landmarks (proximal endpoint, joint, distal endpoint) and computed using the standard dot-product formula for the angle between two vectors:

$$\cos \theta = \frac{(p_1 - p_2) \cdot (p_3 - p_2)}{\|p_1 - p_2\| \|p_3 - p_2\|}$$

where p_2 is the joint position and p_1, p_3 are the adjacent landmarks.

³<https://ai.google.dev/edge/mediapipe/solutions/guide>

The following joint angle features were derived (for both left and right sides where applicable):

- Elbow flexion/extension: the angle between upper arm and forearm (shoulder–elbow–wrist).
- Shoulder angle: the angle between torso and upper arm (hip–shoulder–elbow).
- Wrist flexion/extension: the angle between forearm and hand (elbow–wrist–index finger).
- Wrist radial/ulnar deviation: the wrist angle in the coronal plane (elbow–wrist–pinky finger).
- Thumb–index span: the angle between thumb and index finger (wrist–thumb–index finger).

To ensure data quality, only frames in which all relevant landmarks had nontrivial visibility (MediaPipe visibility > 0.01) were retained, where the visibility score reflects the likelihood that a landmark is present within the image. The resulting pose features comprise continuous time series segments of joint angles (in degrees) describing upper-limb movement over time. These angle time series segments were then resampled to 2 Hz (`scipy.signal.resample`) to match the time steps of the neural and accelerometer features, ensuring all modalities were temporally aligned prior to analysis.

3.4.9 Encoding Model Construction

Encoding models were used to quantify how behavioral variables relate to neural activity. Regularized linear regression was fit to predict multivariate neural measurements from behavioral features. Neural measurements consisted of log-power spectra from three recording sites—the STN, precentral cortex, and postcentral cortex. At each time point, neural data were represented as a vector of size 3 channels $\times$ 251 frequency bins spanning 0–250 Hz. All models were estimated and evaluated separately for each patient and brain side, using the contralateral hand as appropriate. The two main analyses conducted using these data were a *multimodality analysis* and a *stimulation-level analysis*.

In the *multimodality analysis*, we fit separate encoding models using five alternative feature sets:

1. mean watch acceleration (a single scalar per time point, computed as the mean across all log-power accelerometer channels),
2. full watch-acceleration spectrogram,
3. one-hot task labels,

4. pose-based joint-angle features, and
5. a multimodal set combining task labels, watch acceleration, and pose features.

This analysis was restricted to time points from preferred- and over-stimulation states, and all models used the same architecture and preprocessing pipeline.

In the *stimulation-level analysis*, we instead held the feature space fixed and varied stimulation condition. Using only watch-acceleration features as predictors, we fit separate encoding models for the under-stimulation, preferred-stimulation, and over-stimulation conditions. As in the multimodality analysis, all models used an identical architecture and preprocessing pipeline.

3.4.10 Feature Preprocessing

For each recording day, watch-acceleration features (full spectrogram and/or mean power), pose features, and neural features were standardized using a separate z-score transform per day. Task-label features were centered by subtracting the per-day mean of each one-hot column. This per-day normalization ensured that slow calibration drifts or day-to-day behavioral/neural differences could not dominate the regression.

Temporal relationships between predictors and neural activity were modeled using time-lagged features. The predictor matrix was expanded by applying discrete delays of $-4, -2, 0, +2$, and $+4$ samples relative to the neural time axis. At the 2 Hz sampling rate, these delays correspond to $-2, -1, 0, +1$, and $+2$ s. Time-delayed copies of the predictors were concatenated along the feature dimension, allowing each model to learn a linear temporal kernel spanning this ± 2 s window.

3.4.11 Model Estimation

Regularized linear regression was used to map the delayed predictors onto the multi-variate neural response. Ridge regression was chosen because it handles high-dimensional feature spaces (many delays $\times$ many features) and correlated predictors robustly. A multi-output ridge regression (L_2 -regularized linear model) was employed, implemented using the `himalaya` library with GPU acceleration (La Tour et al., 2022). The ridge penalty λ was selected by internal cross-validation over a logarithmically spaced grid from 10^{-6} to 10^6 .

Cross-validation followed a leave-one-run-out scheme at the level of recording days. For each fold, all samples from one day served as the test set, and samples from the remaining days served as the training set. Within the training set, a 5-fold internal cross-validation was used to select the optimal ridge penalty. The model was then refit on the full training data using this optimal λ , and predictions were generated for the held-out day. This procedure was repeated so that each recording day served once as the test fold. Patient 01 was missing left watch acceleration data for one preferred-stimulation block;

consequently, all stimulation-level analyses in the right hemisphere of this patient are based on 5-fold cross-validation.

3.4.12 Model Evaluation

For each fold, performance metrics were computed separately for every neural output dimension (channel $\times$ frequency bin). Pearson correlation r between predicted and observed neural activity was computed on the held-out day for each channel–frequency bin. Correlation values were then averaged across test folds to yield an overall performance estimate per channel–frequency bin for each model and condition. All reported metrics are based exclusively on test data that were not used for training or hyperparameter selection, providing an unbiased estimate of how well the behavioral features can predict neural spectral power in this dataset.

3.4.13 Statistical Analysis

We evaluated how well behavioral features predicted neural spectral activity using regularized encoding models, and quantified model performance using Pearson correlation (r) between predicted and observed neural power on held-out test data. Correlations were computed independently for each channel and cross-validation fold. Channels were defined as the Cartesian product of frequency bin (0–250 Hz; 251 bins) and recording site (STN, precentral gyrus, postcentral gyrus), yielding 753 channels per hemisphere. Patient-specific stimulation artifact bands were excluded from all analyses (Patient 01: 120–140 Hz and 240–250 Hz; Patient 02: 140–160 Hz).

To enable parametric testing, correlations were Fisher z -transformed:

$$z = \text{atanh}(r),$$

and all statistical tests were conducted on z values. Effect sizes are reported in correlation units by back-transforming the mean z : $r = \tanh(\bar{z})$. All statistical analyses were implemented in Python using SciPy and statsmodels. Multiple comparisons were corrected using the Benjamini–Hochberg False Discovery Rate (FDR) procedure at $\alpha = 0.05$.

To assess the benefit of multimodal feature sets, we compared a full multimodal encoding model (“accel + label + pose”) to single-modality models (“mean accel only”, “accel only”, “pose only”, and “task only”). For each patient, hemisphere, and region of interest (ROI), we summarized model performance by averaging z -transformed correlations across non-artifact frequency bins, yielding one mean z per ROI and fold. The best single-modality model was identified as the one with the highest average r across folds. We then compared the full model against this best single-modality model using two-sided paired t -tests across folds. For each contrast, we report the difference in performance ($\Delta r = r_{\text{full}} - r_{\text{single}}$). All pairwise comparisons between single-modality models were similarly tested using paired t -tests on foldwise z values.

To assess how DBS amplitude modulated neural encoding of movement, we compared accelerometry-only encoding models trained separately under three stimulation conditions: under-stimulation, clinically preferred stimulation, and over-stimulation. For each patient, hemisphere, ROI, and frequency band (beta: 13–30 Hz; gamma: 60–90 Hz), we averaged z -transformed correlations across channels within that band and region to produce one summary z per fold and condition. Encoding performance was then compared between stimulation levels using two-sided paired t -tests across matched folds: preferred vs. under-stimulation, and over vs. under-stimulation. For each contrast, we report the difference in performance. All pairwise comparisons between stimulation-level models were similarly tested using paired t -tests on foldwise z values.

3.4.13.1 Model weight adjustments and principal component analysis

For each patient, hemisphere, and region, the full watch–acceleration encoding model produced, for every neural frequency channel, a vector of regression weights mapping accelerometer features onto neural spectral power. Because models were fit separately on multiple cross-validation folds, each neural frequency channel was associated with one weight vector per fold. To obtain a single, stable estimate of the watch–acceleration weights for each region, we averaged the regression coefficients across folds. This procedure yielded, for each region, a frequency-by-feature weight matrix in which each row corresponded to one neural frequency channel and each column to one watch–acceleration feature.

To optimize prediction accuracy, the regularization strength was tuned separately for each neural frequency channel. Differences in signal-to-noise ratio across channels can therefore lead to different optimal regularization hyperparameters and, in turn, to differences in the overall scale of the estimated weights that do not necessarily reflect differences in functional tuning. To reduce this confound, we re-scaled the weights at each neural frequency so that their overall magnitude matched the observed prediction accuracy at that frequency. Specifically, for each neural frequency channel we computed the Euclidean (L2) norm of its weight vector, then applied a scalar normalization factor so that the norm of the re-scaled weight vector equaled the prediction correlation for that channel. Normalized weight matrices were then cropped to the patient-specific analysis ranges (0–120 Hz for Patient 01, 0–140 Hz for Patient 02), excluding high-frequency ranges dominated by artifacts.

To summarize shared structure in the watch–acceleration weights across neural frequencies in precentral and postcentral cortex, we applied principal component analysis (PCA) to the cropped, normalized weight matrix from each region. PCA was performed on the frequency-by-feature matrix, treating neural frequency channels as observations and accelerometer features as variables. Projecting the neural spectral features over time onto the first principal component (PC 1) yielded a single loading value for each time

sample.

3.5 Results

3.5.1 Training Multimodal Encoding Models

To quantify how behavioral variables relate to neural activity, we trained regularized linear regression models to predict multivariate neural signals from behavioral features. Neural activity was summarized as log-power spectra from three recording sites: the STN, precentral gyrus, and postcentral gyrus. At each time point, neural data were represented as a vector of size 3 channels $\times$ 251 frequency bins spanning 0–250 Hz. All models were fit and evaluated separately for each patient and hemisphere, using behavioral features from the contralateral body side.

In the *multimodality analysis*, we compared five encoding models differing only in predictor set:

1. mean watch acceleration (a single scalar per time point, computed as the mean across all log-power accelerometer channels),
2. full watch-acceleration spectrogram,
3. one-hot task labels,
4. pose-based joint-angle features, and
5. a multimodal set combining task labels, watch acceleration, and pose features.

Mean acceleration provides a compact summary of overall movement magnitude, whereas the full spectrogram and pose features capture temporal and kinematic structure at finer resolution. Example time series for behavioral and neural spectral features from one recording day in each patient are shown in Figures 3.1 and 3.2. Qualitatively, behavior-related modulations in the neural spectra are evident in Patient 01 but largely absent in Patient 02. The encoding models quantify these relationships and assess their statistical reliability.

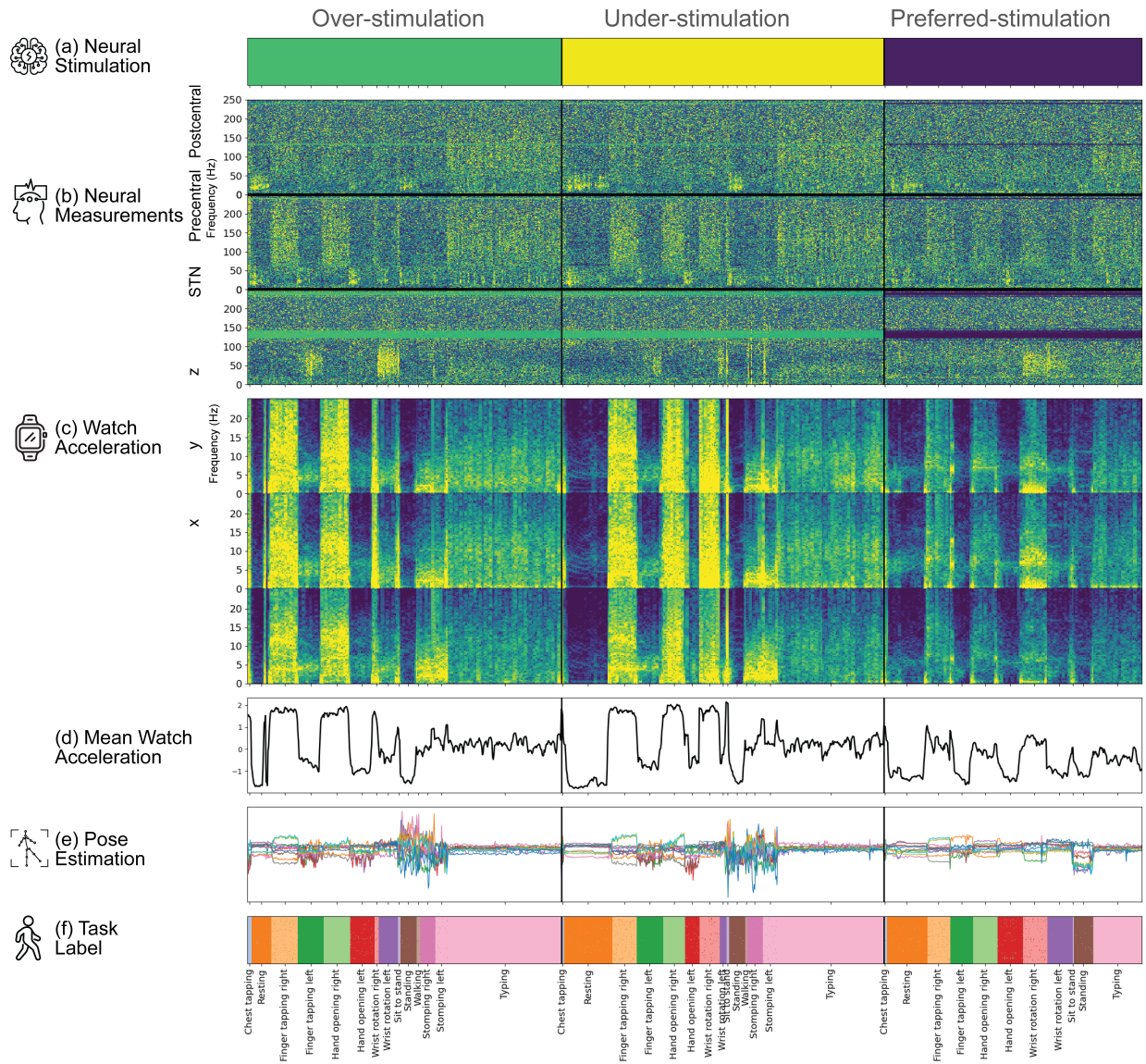


Figure 3.1: Multimodal recordings across stimulation conditions, Patient 01. Overview of simultaneously acquired neural and behavioral signals from one recording day in Patient 01 (left hemisphere), during structured motor tasks under three stimulation states: over-stimulation (left), under-stimulation (center), and preferred-stimulation (right). Clear movement-related modulations are evident in the neural spectra features. **(a) Stimulation state.** Continuous annotation of stimulation level over time, indicating epochs of over-, under-, and preferred-stimulation for a fixed pulse frequency of 130 Hz. **(b) Neural spectral features.** Log-power spectrograms of neural measurements recorded from three sites: STN, precentral gyrus, and postcentral gyrus. Each spectrogram spans 0–250 Hz, is z-scored within the day, and clipped at the 5th and 95th percentiles for visualization. Patient-specific stimulation artifact bands (120–140 Hz and 240–250 Hz for Patient 01) are visible in the raw spectra but were excluded from all subsequent encoding analyses. **(c) Watch-acceleration spectral features.** Log-power spectrograms of wrist acceleration for the three accelerometer axes (X, Y, Z). Each spectrogram spans 0–25 Hz, is z-scored within the day, and clipped at the 5th and 95th percentiles for visualization. **(d) Mean watch acceleration.** Mean watch acceleration, computed as the mean across all log-power watch-acceleration spectrograms. This scalar feature provides a compact summary of overall movement magnitude. **(e) Pose features.** Upper-limb pose features plotted as joint-angle trajectories derived from video-based 3D pose estimation, illustrating arm and hand kinematics across the session (e.g., shoulder, elbow, wrist, and finger joints). **(f) Task labels.** Time-aligned task annotations indicating which instructed motor task is being performed over time.

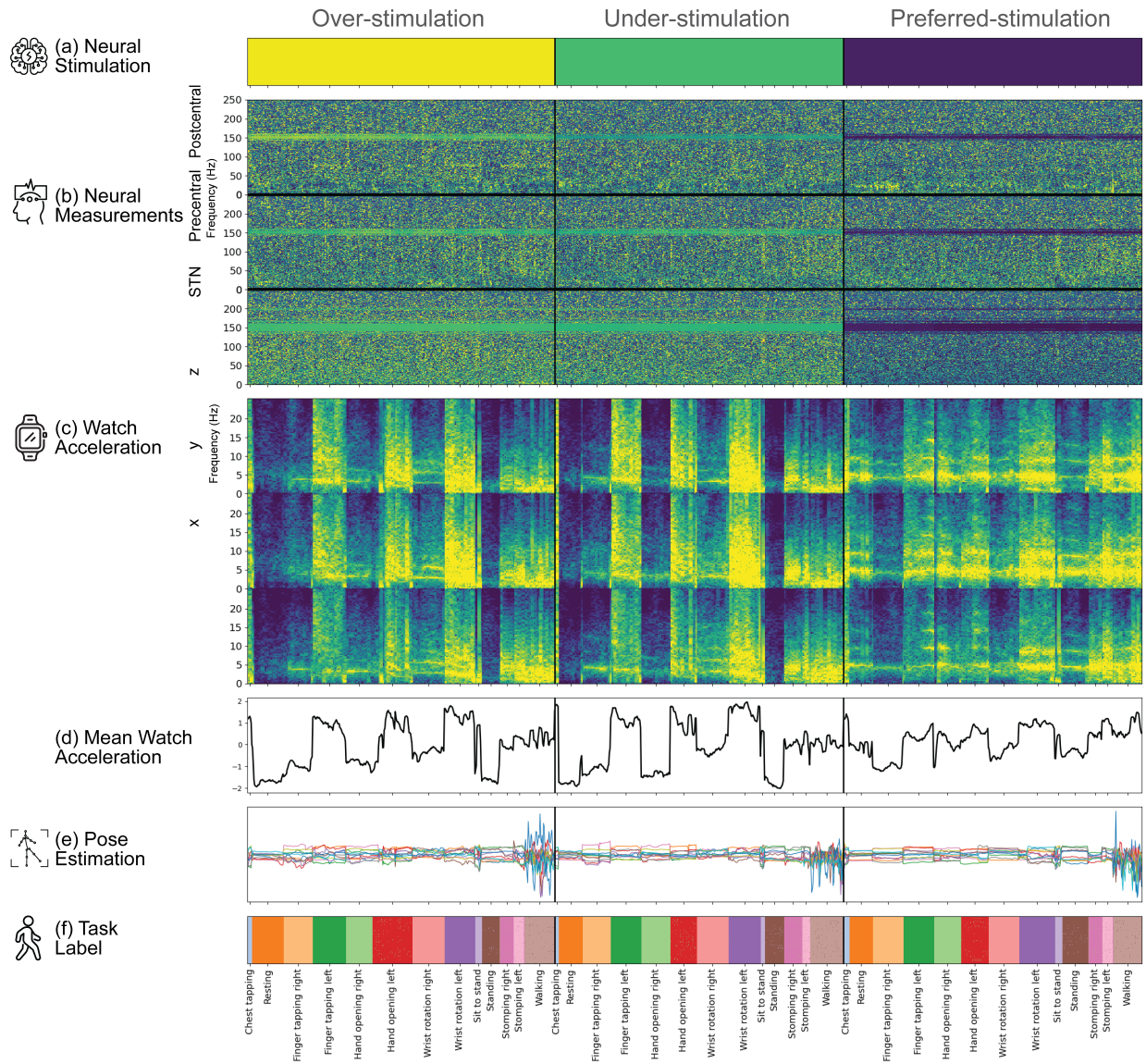


Figure 3.2: Multimodal recordings across stimulation conditions, Patient 02.

Overview of simultaneously acquired neural and behavioral signals from one recording day in Patient 02 (left hemisphere), during structured motor tasks under three stimulation states: over-stimulation (left), under-stimulation (center), and preferred-stimulation (right). Movement-related modulations are not evident in the neural spectra features as are in those of Patient 01. **(a) Stimulation state.** Continuous annotation of stimulation level over time, indicating epochs of over-, under-, and preferred-stimulation for a fixed pulse frequency of 150 Hz. **(b) Neural spectral features.** Log-power spectrograms of neural measurements recorded from three sites: STN, precentral gyrus, and postcentral gyrus. Each spectrogram spans 0–250 Hz, is z-scored within the day, and clipped at the 5th and 95th percentiles for visualization. Patient-specific stimulation artifact bands (140–160 Hz for Patient 02) are visible in the raw spectra but were excluded from all subsequent encoding analyses. **(c) Watch-acceleration spectral features.** Log-power spectrograms of wrist acceleration for the three accelerometer axes (X, Y, Z). Each spectrogram spans 0–25 Hz, is z-scored within the day, and clipped at the 5th and 95th percentiles for visualization. **(d) Mean watch acceleration.** Mean watch acceleration, computed as the mean across all log-power watch-acceleration spectrograms. This scalar feature provides a compact summary of overall movement magnitude. **(e) Pose features.** Upper-limb pose features plotted as joint-angle trajectories derived from video-based 3D pose estimation, illustrating arm and hand kinematics across the session (e.g., shoulder, elbow, wrist, and finger joints). **(f) Task labels.** Time-aligned task annotations indicating which instructed motor task is being performed over time.

3.5.2 Evaluating Encoding Model Prediction Performances

3.5.2.1 Prediction Performances across Patients, Hemispheres, and Regions

Behavioral features robustly predicted neural spectral power across recording sites in both patients, though encoding strength differed markedly between individuals. In Patient 01, the full multimodal model achieved moderate prediction accuracy across artifact-free frequencies (precentral: mean $r = 0.17$; postcentral: mean $r = 0.12$; STN: mean $r = 0.11$). In Patient 02, performance was substantially weaker but preserved the same regional ordering (precentral: mean $r = 0.06$; postcentral: mean $r = 0.05$; STN: mean $r = 0.01$). Fold-wise tests confirmed significantly higher prediction performance in Patient 01 across all regions (all $\Delta r \geq 0.07$, all $q < 0.001$). These individual differences likely reflect variation in signal quality rather than fundamental differences in brain–behavior coupling.

Within each patient, encoding prediction performance was consistently highest in precentral cortex, intermediate in postcentral cortex, and weakest in STN. In Patient 01, precentral predictions significantly exceeded both postcentral ($\Delta r = 0.05$, $q < 0.001$) and

STN ($\Delta r = 0.06$, $q = 0.004$), whereas the postcentral–STN difference was not significant ($q = 0.25$). Patient 02 showed a similar hierarchy: both cortical sites outperformed STN (precentral vs. STN: $\Delta r = 0.06$, $q < 0.001$; postcentral vs. STN: $\Delta r = 0.04$, $q < 0.001$), and precentral modestly exceeded postcentral ($\Delta r = 0.01$, $q = 0.004$). This regional pattern—strongest encoding in motor cortex, weakest in STN—was consistent across patients despite their overall difference in signal strength.

Hemispheric asymmetries were present but opposite in direction between patients. In Patient 01, left precentral cortex showed stronger encoding than right (mean $r = 0.21$ vs. 0.12 , $\Delta r = 0.10$, $q < 0.001$), with a similar but nonsignificant trend in STN ($q = 0.15$); postcentral cortex was symmetric ($q = 0.73$). In Patient 02, the pattern reversed: right hemisphere outperformed left in both precentral ($\Delta r = 0.02$, $q = 0.013$) and postcentral cortex ($\Delta r = 0.01$, $q = 0.013$), while STN remained symmetric ($q = 0.44$). These opposing lateralization patterns likely reflect differences in recording quality rather than consistent functional asymmetries.

The stimulation-entrained gamma band—defined as ± 1 Hz around half the stimulation frequency (65 Hz for Patient 01, 75 Hz for Patient 02)—showed particularly strong encoding. In Patient 01, entrained-band correlations were robustly above zero across all regions (e.g., left precentral: mean $r = 0.41$; left STN: mean $r = 0.38$; all $q < 0.001$) and substantially exceeded full-band averages ($\Delta r = 0.19$ – 0.25 , all $q < 0.05$). In Patient 02, cortical sites similarly showed elevated entrained-band encoding relative to full-band means (e.g., right postcentral: mean $r = 0.41$, $\Delta r = 0.36$, $q < 0.001$), but STN showed no reliable entrained-band prediction (both hemispheres $q > 0.29$). These results indicate that stimulation-entrained gamma carries especially strong movement-related information in sensorimotor cortex for both patients and in STN for Patient 01.

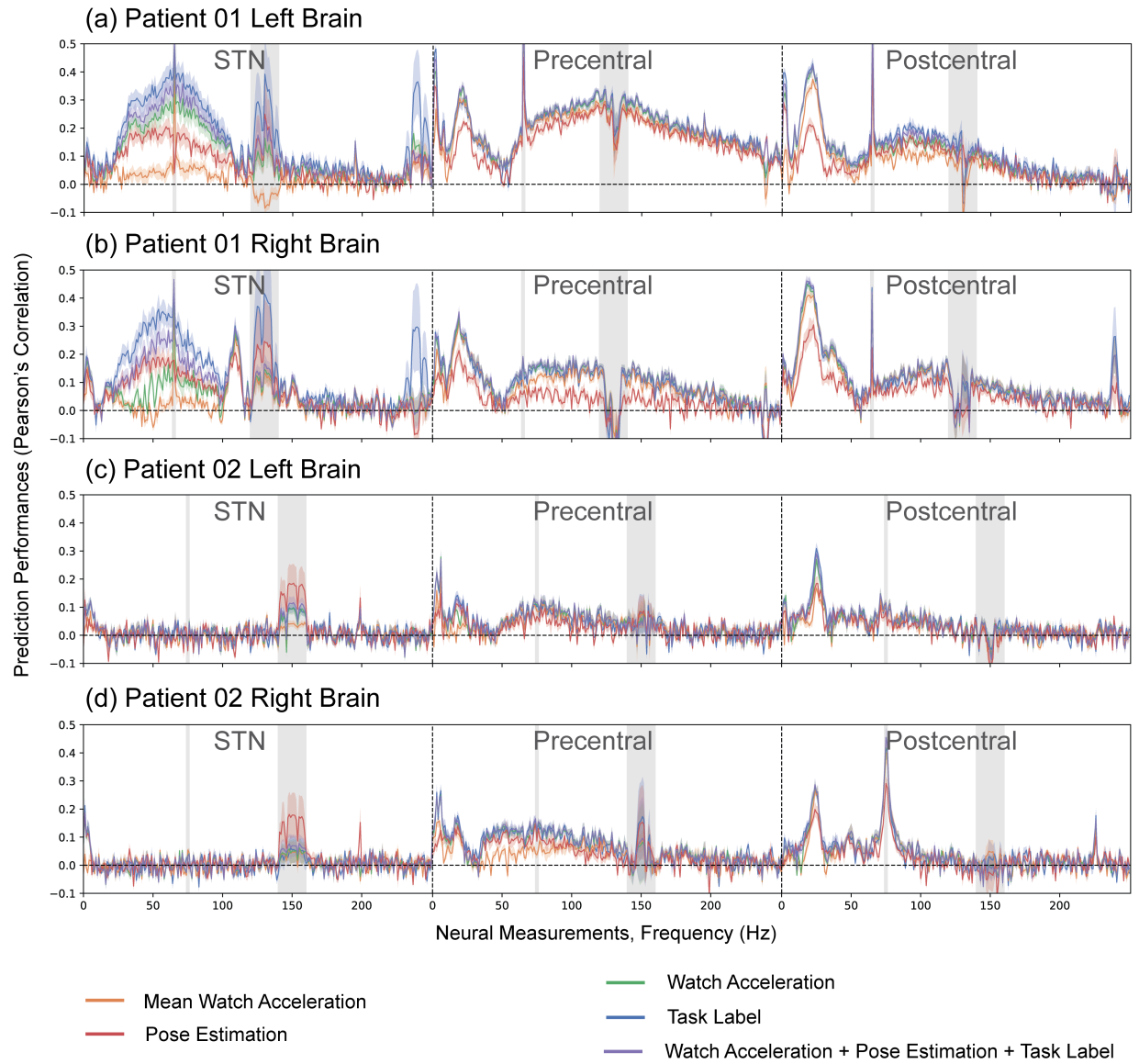


Figure 3.3: Behavioral features predict neural spectral activity at sensing contacts (a–d) Cross-validated prediction performance (Pearson’s r) for unimodal and multimodal encoding models plotted against neural frequency, shown separately for each patient and hemisphere (rows) and for STN, precentral cortex, and postcentral cortex (segments separated by vertical dashed lines). Colored lines show mean prediction accuracy across cross-validation folds for each model (orange: mean watch acceleration; red: pose estimation; green: full watch acceleration spectrogram; blue: task labels; purple: full multimodal model combining all three), with shaded bands indicating the standard error across folds. Thick gray-shaded regions mark stimulation artifact frequencies that were excluded from all statistical analyses; narrow peaks near half the stimulation frequency (65 Hz in Patient 01, 75 Hz in Patient 02) correspond to the stimulation-entrained gamma band. Across panels, the full watch acceleration spectrogram and task labels produce the largest peaks and highest overall correlations, whereas mean watch acceleration is consistently the weakest predictor; the multimodal model provides modest additional gains over the best unimodal model, particularly in sensorimotor cortex. Prediction performance is markedly higher in Patient 01 than in Patient 02 across all regions. These patterns indicate that behaviorally driven variance in neural activity is robustly expressed at the sensing contacts of the neurostimulator system, and that rich temporal structure in watch acceleration and task timing carries most of the behaviorally relevant predictive information, beyond what is captured by mean acceleration as a coarse movement summary. They also reveal substantial individual differences in overall prediction strength, likely reflecting differences in signal-to-noise ratio and recording quality between patients.

3.5.2.2 Relative Contribution of Behavioral Features

Comparison of unimodal encoding models revealed a consistent hierarchy: temporally rich predictors (full acceleration spectrogram, task labels) substantially outperformed coarse movement summaries (mean acceleration), with pose-based features intermediate. Across both patients and all recording sites, mean acceleration was consistently the weakest unimodal predictor, whereas full acceleration spectrogram and task labels provided the strongest predictions. In Patient 01, full acceleration outperformed mean acceleration by $\Delta r = 0.02$ – 0.08 across cortical and STN sites (all $q < 0.001$); task labels showed similar advantages over mean acceleration ($\Delta r = 0.05$ – 0.13 , all $q < 0.001$). In Patient 02, effect sizes were smaller but the same ordering held: full acceleration exceeded mean acceleration by $\Delta r = 0.01$ – 0.03 in cortex (all $q < 0.01$), and task labels similarly outperformed mean acceleration across regions. This consistent pattern indicates that fine-grained temporal and categorical structure in behavior is necessary for capturing movement-related variance in sensed neural signals.

Combining behavioral modalities yielded modest additional gains over the best

unimodal predictor at select cortical sites. In Patient 01, the multimodal model (full acceleration + task labels + pose) significantly outperformed the best unimodal baseline in left precentral cortex ($\Delta r = 0.005$, $q < 0.001$) and right postcentral cortex ($\Delta r = 0.003$, $q < 0.01$), with no reliable advantage elsewhere. Patient 02 showed a similar pattern: small multimodal gains in left precentral ($\Delta r = 0.002$, $q < 0.05$) and right postcentral cortex ($\Delta r = 0.003$, $q < 0.01$), but not in STN or other cortical sites. These improvements, though statistically reliable, were quantitatively modest—typically an order of magnitude smaller than the gap between rich and coarse unimodal predictors.

These results indicate that local field potentials and electrocorticographic signals recorded from implanted electrodes encode the spectral and categorical structure of ongoing motor behavior, not simply movement presence or magnitude. This structure is captured by temporally resolved and task-structured predictors but inaccessible to scalar summaries. Multimodal integration provides minimal additional benefit once fine-grained behavioral structure is represented. Rest-versus-movement contrasts prevalent in prior studies thus appear insufficient for characterizing the full extent of brain–behavior coupling available in chronically sensed signals.

3.5.3 Stimulation Amplitude Modulates Encoding Model Performances

To assess how stimulation amplitude affects brain–behavior coupling, we fit watch-acceleration encoding models separately for under-, preferred-, and over-stimulation conditions and compared prediction accuracy across conditions using fold-wise paired tests (FDR-corrected; Figure 3.4, 3.5).

In Patient 01, elevated stimulation amplitude significantly increased encoding performance across the sensorimotor network. Full-spectrum prediction accuracy was higher during both preferred and over-stimulation than during under-stimulation in left STN and bilateral sensorimotor cortex ($\Delta r = 0.04$ – 0.09 , all $q < 0.02$). Right STN showed a similar pattern, though only the over versus under contrast reached significance ($\Delta r = 0.03$, $q < 0.05$). Band-resolved analyses revealed that cortical enhancement spanned both beta (13–30 Hz) and gamma (60–90 Hz) frequencies, with significant increases at all cortical sites for both stimulation contrasts ($\Delta r = 0.03$ – 0.13 , all $q < 0.05$). STN effects were gamma-selective: over-stimulation increased gamma-band encoding bilaterally (left: $\Delta r = 0.22$; right: $\Delta r = 0.08$; both $q < 0.05$), and preferred stimulation increased left STN gamma ($\Delta r = 0.13$, $q < 0.05$) but not right ($q > 0.05$). STN beta showed no enhancement; right STN beta was modestly reduced at preferred stimulation ($\Delta r = -0.05$, $q < 0.05$).

In Patient 02, stimulation effects were weaker and spatially restricted. Full-spectrum contrasts were not significant in STN or left cortex ($q > 0.07$). The only significant broadband effect was increased prediction accuracy in right postcentral cortex during over-stimulation ($\Delta r = 0.02$, $q < 0.01$). Band-resolved analyses localized this enhancement to gamma: both preferred and over-stimulation increased right postcentral gamma encoding

($\Delta r = 0.06$ and 0.09 , respectively; both $q < 0.05$), with no significant beta changes. Left precentral cortex showed reduced beta-band encoding at preferred stimulation ($\Delta r = -0.05$, $q < 0.05$), opposite to the pattern in Patient 01. No other contrasts survived correction.

The divergent patterns between patients, widespread enhancement in Patient 01 versus focal effects in Patient 02, parallel their differences in baseline encoding strength, suggesting that stimulation-related modulation is more readily detected when underlying signal quality is higher.

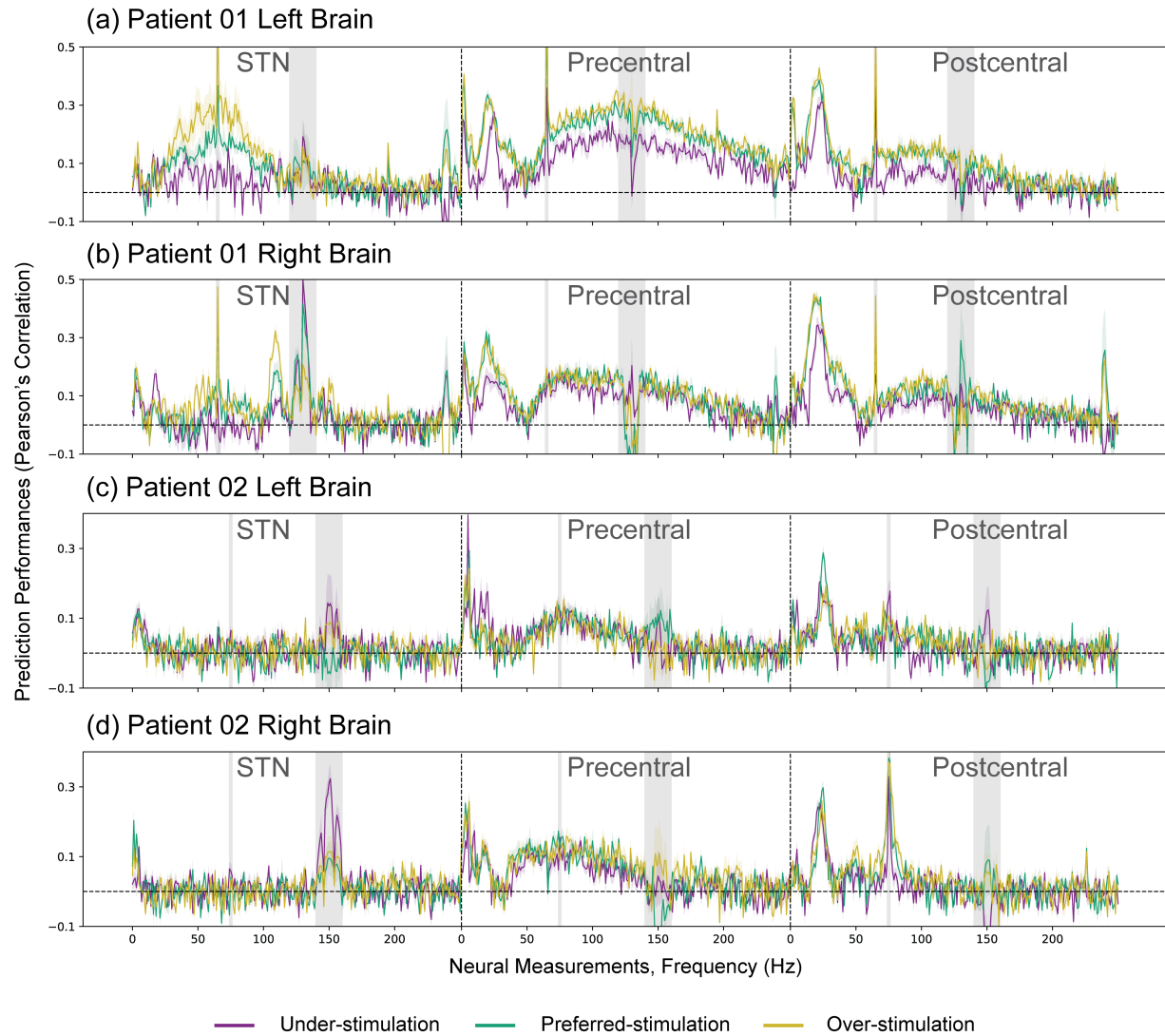


Figure 3.4: Stimulation amplitude modulates encoding model performance (a–d) Cross-validated prediction performance (Pearson’s r) of the watch acceleration encoding model across neural frequency, shown separately for each patient and hemisphere (rows) and for STN, precentral cortex, and postcentral cortex (segments separated by vertical dashed lines). Colored lines show mean prediction accuracy across cross-validation folds for each stimulation level (purple: under-stimulation; green: preferred-stimulation; yellow: over-stimulation), with shaded bands indicating the standard error across folds. Gray-shaded regions mark stimulation artifact frequencies that were excluded from statistical analyses; narrow peaks near half the stimulation frequency correspond to stimulation-entrained gamma (65 Hz in Patient 01, 75 Hz in Patient 02). In Patient 01, preferred and over-stimulation markedly increase prediction performance throughout STN and sensorimotor cortex, whereas in Patient 02 stimulation-related changes are weaker and increase in prediction performance is largely confined to gamma band in right postcentral cortex.

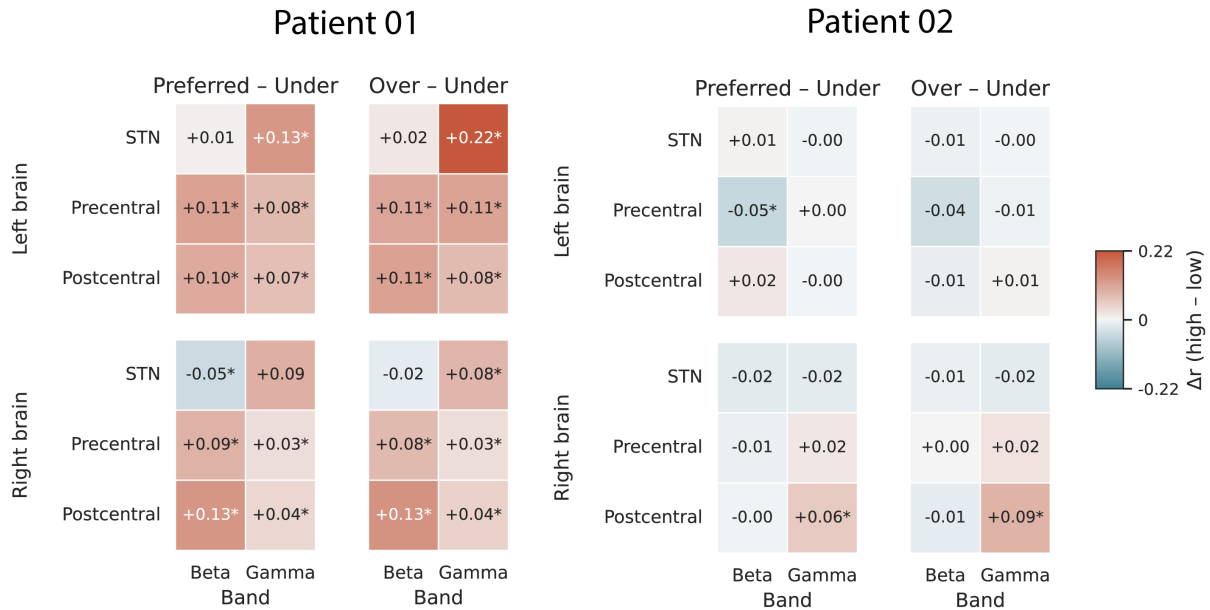


Figure 3.5: Stimulation amplitude modulates encoding model performance in canonical oscillation bands across regions and patients Heatmaps show changes in cross-validated prediction performance (Pearson’s r) of the watch-acceleration encoding model for beta (13–30 Hz) and gamma (60–90 Hz) bands across stimulation levels. Columns indicate contrasts between stimulation conditions (preferred – under and over – under), and rows indicate region (STN, precentral, postcentral) and hemisphere (left brain, right brain), plotted separately for **Patient 01** (left panel) and **Patient 02** (right panel). Each cell reports the mean change in correlation (Δr , high – low stimulation), with color encoding the magnitude and direction of the effect (blue: decreased, red: increased prediction performance); asterisks mark FDR-corrected significant effects ($q < 0.05$). Patient 01 shows widespread stimulation-related enhancement of beta- and gamma-band prediction performance in precentral and postcentral cortex and gamma-band prediction performance in STN. In contrast, Patient 02 exhibits only localized effects, with increased gamma-band prediction performance confined primarily to right postcentral cortex.

3.5.4 Stimulation Amplifies a Movement-Locked Cortical Component

To understand why stimulation enhanced encoding-model prediction performance in Patient 01, we asked whether higher stimulation amplifies a specific movement-related pattern of neural frequency modulation. Inspection of the stimulation-level spectra (Figure 3.1) suggested that preferred and over-stimulation produced larger beta and gamma peaks in precentral and postcentral cortex than under-stimulation, consistent with stronger

movement-related modulation at higher stimulation levels. To test this hypothesis, we performed principal component analysis (PCA) on the watch acceleration weight matrices for precentral and postcentral cortex. For each region, we arranged weights as a neural-frequency by accelerometry-frequency matrix and extracted the first principal component (PC 1; Figure 3.6a,b). PC 1 exhibited a characteristic spectral profile: positive loadings in cortical gamma—including a pronounced peak at the stimulation-entrained subharmonic—negative loadings in beta, and additional positive loadings at delta/alpha frequencies. This multiband structure is consistent with movement-related beta desynchronization coupled with gamma enhancement, a signature of pro-kinetic cortical states.

We quantified PC 1 dynamics using two metrics: root-mean-square (RMS) amplitude as a measure of component expression, and the correlation between the PC 1 projection and mean watch acceleration as a measure of movement tracking. In the left hemisphere, both metrics increased with stimulation amplitude. PC 1 expression was significantly higher during preferred- and over-stimulation than under-stimulation in precentral cortex (RMS: 15.42 vs. 16.41 and 16.49; $p \leq 0.003$), with no difference between preferred and over ($p = 0.69$). Postcentral cortex showed a significant under-to-over increase (RMS: 11.46 vs. 12.21; $p = 0.005$) and a marginal under-to-preferred trend ($p = 0.050$). Movement tracking, quantified as the magnitude of the correlation between PC 1 and acceleration, improved monotonically with stimulation: in precentral cortex, mean $|r|$ rose from 0.459 (under) to 0.569 (preferred) to 0.631 (over), with all pairwise comparisons significant ($p \leq 0.048$). Postcentral cortex exhibited an analogous pattern (mean $|r|$: 0.454 to 0.538 to 0.605), with significant under-to-preferred ($p = 0.031$) and under-to-over effects ($p = 0.001$), though the preferred-to-over increase did not reach significance ($p = 0.097$).

The right hemisphere showed improved movement tracking but weaker stimulation dependence of component expression. Correlations between PC 1 and acceleration increased substantially from under-stimulation to both preferred- and over-stimulation in precentral cortex (mean r : 0.414 to 0.551 to 0.568; under vs. preferred/over: $p \leq 0.001$) and postcentral cortex (mean r : 0.383 to 0.593 to 0.595; $p \leq 0.006$), but preferred and over conditions did not differ in either region ($p \geq 0.16$). In contrast, PC 1 expression showed limited stimulation dependence: precentral RMS did not differ across conditions (all $p \geq 0.13$), and postcentral RMS increased only from under- to preferred-stimulation (13.16 vs. 13.96; $p = 0.003$), with no further change at over-stimulation.

These results support the hypothesis that higher stimulation selectively amplifies a movement-related cortical component characterized by coordinated modulation across beta, gamma, and alpha/delta bands. The clearest amplification occurred in the left hemisphere, where both magnitude and movement-tracking fidelity of PC 1 increased with stimulation. These findings provide a mechanistic link between stimulation level and improved encoding performance, consistent with stimulation-dependent facilitation of movement-related information flow through sensorimotor cortex.

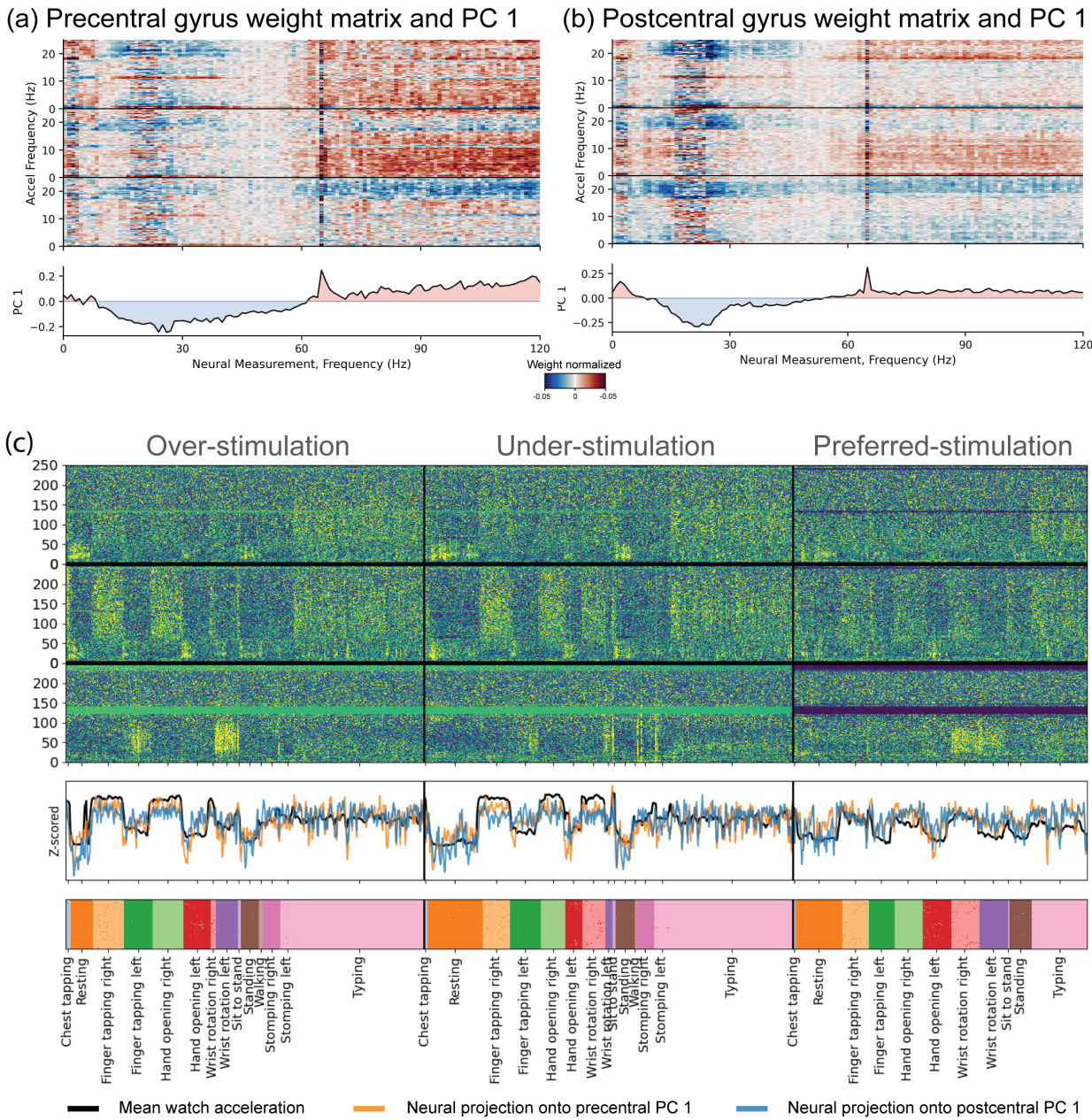


Figure 3.6: Stimulation amplifies a movement-related cortical component (a,b) Encoding model weight matrices mapping watch-acceleration features (y-axis: accelerometer frequency) onto neural spectral power (x-axis: neural frequency) for precentral (a) and postcentral (b) cortex in Patient 01, left hemisphere. Each row of the weight matrix corresponds to one accelerometer frequency band; each column corresponds to one neural frequency bin. Color indicates normalized regression weight (red: positive; blue: negative). Below each matrix, the first principal component (PC 1) summarizes the dominant structure of the weights across neural frequencies. PC 1 exhibits negative loadings in the beta band ($\sim 13\text{--}30$ Hz), positive loadings at lower (delta/alpha) frequencies, and a prominent positive peak at the stimulation-entrained gamma frequency (~ 65 Hz, half the 130 Hz stimulation frequency). (c) Example recording day illustrating how PC 1 tracks movement across stimulation conditions. Top three rows: neural spectrograms from STN, precentral, and postcentral cortex (segments separated by vertical dashed lines) during over-stimulation (left), under-stimulation (center), and preferred-stimulation (right). Fourth row: time-aligned behavioral and neural signals, showing mean watch acceleration (black), neural projection onto precentral PC 1 (orange), and neural projection onto postcentral PC 1 (blue), all z-scored. Bottom row: task labels indicating the motor task performed at each time point. During preferred- and over-stimulation, the cortical PC 1 projections more closely track task-related fluctuations in mean acceleration, whereas this coupling is attenuated during under-stimulation.

3.6 Discussion

This study leveraged a home-deployed multimodal recording platform to investigate how DBS amplitude modulates the encoding of movement-related information in neural activity recorded from the subthalamic nucleus and sensorimotor cortex in patients with Parkinson’s disease. Our findings demonstrate that behavioral features robustly predict neural spectral power across recording sites, that temporally rich representations substantially outperform coarse movement summaries, and that higher stimulation amplitudes enhance brain–behavior coupling in a manner consistent with informational-lesion accounts of DBS.

Across both participants, hemispheres, and recording sites, encoding models trained on behavioral features successfully predicted neural spectral activity. This finding establishes that behaviorally driven variance is clearly expressed at the sensing contacts of bidirectional neurostimulators, providing an empirical foundation for adaptive DBS strategies that incorporate behavioral context. Critically, the choice of behavioral representation mattered substantially: the full watch-acceleration spectrogram and task labels consistently outperformed mean acceleration, indicating that fine-grained temporal and

categorical structure in movement is essential for capturing neural variance. This result has practical implications for adaptive DBS design, suggesting that simple on/off or gross movement state representations may fail to capture much of the task- and movement-related modulation available in sensed neural signals.

In Patient 01, who exhibited stronger overall encoding, preferred and over-stimulation markedly enhanced prediction performance throughout STN and sensorimotor cortex. This effect was observed in both hemispheres, each driven by an independent neurostimulator and modeled with a separate encoding model, providing within-subject replication across two independently recorded and analyzed neural systems. Principal component analysis of the acceleration encoding weights revealed that the dominant component exhibited negative beta-band loadings coupled with positive loadings in alpha/delta and gamma bands, including a prominent peak at stimulation-entrained gamma. At higher stimulation amplitudes, this component tracked movement more faithfully, consistent with stimulation-dependent amplification of movement-related neural modulation.

Our findings accord with “information lesion” accounts of DBS mechanism (Grill et al., 2004; Rubin and Terman, 2004). Under this framework, the therapeutic effect of DBS arises not from simple suppression or enhancement of activity in particular frequency bands, but from restoration of the flow of movement-related information through basal ganglia–cortical circuits. The pathological state in Parkinson’s disease is characterized by excessive beta synchronization that disrupts the transmission of behaviorally relevant signals (Brown, 2003; Brittain and Brown, 2014). DBS, rather than merely reducing beta power, may act by regularizing network dynamics in a manner that permits movement-related information to propagate more effectively.

These findings extend prior work demonstrating that DBS modulates beta and gamma power by showing that the functional consequence of this modulation is enhanced brain–behavior coupling. To our knowledge, this stimulation-dependent amplification of a movement-related oscillatory component has not previously been demonstrated in human patients.

3.6.1 Limitations

Several limitations of this study warrant consideration.

The most significant limitation is the small sample size of two participants. This constraint reflects the practical challenges inherent in chronic human intracranial recording studies: the investigational Summit RC+S neurostimulator is implanted in a limited number of patients at specialized centers, the multimodal home recording platform requires substantial technical infrastructure and participant training, and the protocol demands considerable participant time and commitment across multiple recording days. These factors severely constrain the feasible sample size for studies of this nature.

However, several features of our design partially mitigate this limitation. First, each

participant contributed data from both hemispheres, with each hemisphere driven by an independent neurostimulator and analyzed with a separate encoding model. This within-subject replication across four independently recorded neural systems (two hemispheres $\times$ two participants) provides stronger evidence than would a single recording per participant. Second, cross-validation was performed at the level of recording days, with six-fold leave-one-day-out evaluation, ensuring that reported performance metrics reflect generalization to held-out data rather than overfitting to the training set. Third, statistical tests were conducted on fold-wise metrics with appropriate correction for multiple comparisons, providing valid inference within the constraints of the available data.

Nevertheless, the extent to which these findings generalize to the broader population of patients with Parkinson's disease remains uncertain. Individual differences in disease phenotype, electrode placement, stimulation parameters, and neural signal quality may all influence the relationship between stimulation, behavior, and neural activity. Future studies with larger cohorts will be essential to characterize the heterogeneity of these effects and identify patient-specific factors that predict stimulation-related enhancement of brain-behavior coupling.

A second limitation concerns the marked difference in encoding model performance between the two participants. Patient 01 exhibited substantially higher prediction accuracy across all regions, hemispheres, and behavioral feature sets compared to Patient 02. This difference likely reflects variation in signal-to-noise ratio and recording quality rather than fundamental differences in the underlying neural computations. Factors that may contribute to such variation include electrode impedance, precise contact placement relative to regions of interest, and extent of local tissue reaction.

The lower signal quality in Patient 02 limited our ability to detect stimulation modulation effects in this individual. While Patient 01 showed robust stimulation-dependent enhancement of encoding model performance across bilateral STN and sensorimotor cortex, Patient 02 exhibited only localized effects, with increased gamma-band prediction performance confined primarily to right postcentral cortex. It remains unclear whether this difference reflects a true absence of stimulation modulation effects in Patient 02 or simply inadequate statistical power due to lower baseline signal quality. Improved recording hardware, optimized electrode configurations, and artifact rejection algorithms may enhance signal quality in future studies and permit more sensitive detection of stimulation effects across a broader range of patients.

A third limitation concerns the scope of behavioral sampling. Although our task battery included multiple UPDRS-derived motor tasks and functional activities, it did not exhaustively sample the range of movements and contexts encountered in daily life. The encoding models were trained primarily on structured, repetitive movements performed in a seated position within a home office environment. Whether the observed relationships between stimulation, behavior, and neural activity generalize to more complex, naturalistic behaviors—such as walking in crowded environments, fine manipulation of objects, or

activities requiring cognitive-motor coordination—remains to be established.

3.6.2 Future Directions

These findings suggest several directions for future research. First, future studies should prioritize collection of rich, temporally resolved behavioral data. Our results demonstrate that fine-grained movement structure is essential for capturing behaviorally driven neural variance: the full acceleration spectrogram and task labels substantially outperformed mean acceleration as predictors of neural activity. This finding implies that adaptive DBS systems relying on coarse behavioral summaries may fail to exploit much of the available information linking behavior to neural state. Wearable sensors capable of capturing detailed kinematics, combined with camera-based pose estimation and task context, may enable more precise estimation of behavioral state and, in turn, more effective adaptive stimulation.

Second, the observation that stimulation enhances brain–behavior coupling motivates exploration of stimulation parameters beyond amplitude. Frequency, pulse width, contact configuration, and temporal patterning (e.g., burst stimulation, closed-loop triggered stimulation) may differentially affect the transmission of movement-related information through basal ganglia–cortical circuits. Systematic manipulation of these parameters within the encoding model framework developed here could reveal parameter combinations that optimize brain–behavior coupling while minimizing energy consumption and side effects.

Finally, the encoding model framework developed here could be extended to incorporate decoding models that predict symptom severity or clinical state from neural features. Such bidirectional models—linking behavior to neural activity and neural activity to clinical outcomes—would provide a more complete picture of the causal pathways through which DBS produces therapeutic benefit and could inform the design of adaptive systems that titrate stimulation to maintain optimal brain–behavior coupling.

3.7 Supplementary Materials

3.7.1 Patient Instructions

1. Make sure that the Patient Telemetry Module (PTM), Clinician Telemetry Module (CTM), and both Apple Watches are charged.
2. Put on both Apple Watches and begin streaming Apple Watch data using StrivePD.
3. The first of three blocks will begin. Each of the three blocks will be fifteen minutes long and performed at a different stimulation level. The stimulation levels for each block are mentioned on the schedule at the end of the instructions and the order will

be shuffled each day. Please switch to Group A prior to streaming. Within each block you will:

- Connect the PTM and set the stimulation level for both INS devices (same for both sides; may need to disconnect the CTM if previously streaming).
 - Connect the CTM and begin streaming data for both INS devices (may need to disconnect PTM first).
 - Wait five minutes.
 - Tap the INS on the chest 5 times in any irregular pattern. This will be used for aligning movement data from the RC+S with the video data, so please make the tapping movements large enough to be visible by the cameras.
 - Rest your hands on your lap with palms facing up for 15 seconds. Then count backwards out loud from 100 (hands in same position) for another 15 seconds.
 - Using a single hand at eye-level, perform index-thumb finger tapping with as big and fast of movements as possible for 20 seconds. Repeat with the other hand once finished.
 - Using the whole hand at eye-level, perform hand open-and-close as big and as fast as possible for 20 seconds. Repeat with the other hand once finished.
 - With the hand outstretched, perform hand rotations, all the way clockwise and all the way counterclockwise, as big and as fast as possible for 20 seconds. Repeat with the other hand once finished.
 - With arms crossed—sit to stand, then relax arms by your side and stand still for 15 seconds.
 - Perform walking in view of the camera for 20 seconds. Show foot stamping on each side for 10 seconds while standing.
 - Type the text provided at the end of these instructions. We do not need to see record of the typing, so feel free to do it in any text editor and delete afterwards. This should be completed with a few minutes left in the block. It is completely fine if you find you cannot finish typing the full text with time to spare, so feel free to adjust how much of the text to complete.
 - Relax and do as you please until the next fifteen-minute block. This will be self-paced, but please try to begin the next block close to fifteen minutes after the previous one was started.
4. After the three blocks have been completed, report medication information for the current day.

-
5. Stop streaming data on both the RC+S and Apple Watches. The video recording will stop automatically.
 6. Connect to the PTM and return stimulation to your standard settings on both INS devices.
 7. Take a nice break and have a snack for all your hard work.

CHAPTER 4

Conclusion

This thesis set out to use machine learning not merely to predict neural signals, but to extract neuroscientific insight. Across two seemingly disparate settings—whole-brain fMRI during naturalistic language comprehension and chronic intracranial recordings in Parkinson’s disease—a central theme emerges: the scientific yield of predictive modeling depends jointly on model interpretability and on the quantity and quality of available measurements.

4.1 Summary of Contributions

4.1.1 Chapter 2: Interpretable Brain Encoding Model with Sparse Concept Atoms

Chapter 2 addressed a fundamental limitation of encoding models built on modern language embeddings. Because these embeddings compress many latent semantic factors into a lower-dimensional space, fitted voxel weights conflate concepts that share representational directions—a phenomenon termed superposition. Superposition renders regression coefficients difficult to interpret and, formally, non-identifiable as multiple distinct patterns of conceptual selectivity can produce identical predicted responses. To resolve this ambiguity, I applied sparse dictionary learning to factor dense word embeddings into an overcomplete basis of sparse, interpretable concept atoms, then refit encoding models in this transformed feature space. The resulting Sparse Concept Encoding Model preserved prediction accuracy while substantially improving interpretability: each fitted coefficient now indexes a semantically coherent direction rather than an opaque mixture. This reformulation enabled two analyses that would be ambiguous or intractable in the original dense embedding space. First, by examining weights along distinct concept-atom axes, I disentangled overlapping cortical selectivity for time, space, and number, revealing

convergent representations in intraparietal sulcus and inferior frontal gyrus—a pattern consistent with the shared-magnitude hypothesis. Second, by comparing cortical tuning maps across concept atoms, I exposed systematic structure in the pairwise similarity of distributed semantic maps.

4.1.2 Chapter 3: Deep Brain Stimulation Amplifies Movement-Related Neural Modulation in Parkinson’s Disease

Chapter 3 shifted from controlled laboratory paradigms to patient-centered neuroscience. Using a home-deployed platform integrating sensing-enabled neurostimulators (Medtronic Summit RC+S), bilateral wrist accelerometry, and video-based pose estimation, we recorded synchronized neural and behavioral signals from two individuals with Parkinson’s disease across multiple days, systematically varying both motor task and stimulation amplitude. Encoding models revealed that local field potentials sensed from implanted electrodes carry substantial information about ongoing behavior—information recoverable from wearable accelerometry, task indicators, and pose-derived kinematics alike. Critically, full acceleration spectrograms and categorical task structure substantially outperformed scalar summaries of movement magnitude, demonstrating that fine-grained behavioral dynamics are necessary to account for movement-related variance in the sensed signals. In the participant with stronger overall encoding model performance, elevated stimulation amplitudes increased prediction accuracy throughout subthalamic nucleus and sensorimotor cortex by amplifying a movement-related spectral component: beta-band suppression coupled with enhancement in alpha/delta and gamma bands, including a prominent peak at the stimulation-entrained subharmonic. This pattern aligns with informational-lesion accounts of DBS, in which pathological synchronization impedes movement-related signaling and high-frequency stimulation restores information flow through basal ganglia–cortical circuits. More broadly, these findings support the feasibility of information-rich neural sensing in chronic neuromodulation and motivate closed-loop architectures that incorporate behavioral context rather than treating neural biomarkers as isolated control signals.

4.2 Limitations and Lessons

Taken together, these studies clarify both what machine learning can extract from current neuroscience data and where it falls short.

In Chapter 2, the intrinsic constraints of fMRI—millimeter-scale spatial resolution and hemodynamic blurring on the order of seconds—bound the granularity with which semantic representations can be resolved. The encoding framework identifies *where* conceptual information is expressed across cortex but cannot adjudicate *how* that information is computed at the level of local circuits or single neurons. In Chapter 3, pronounced

individual differences in signal quality limited statistical power: one participant exhibited robust stimulation effects spanning the sensorimotor network, whereas the other showed weaker, more spatially restricted modulation. Whether this asymmetry reflects electrode placement, disease stage, cortical anatomy, or some combination thereof remains an open question. Further, both chapters were constrained to linear regression—a choice that facilitates interpretation of fitted coefficients but forfeits the capacity to capture nonlinear transformations that almost certainly govern neural computation.

If the ultimate objective is to model the brain as the high-dimensional, nonlinear dynamical system it is and to translate such understanding into effective treatments for neurological disease, several methodological gaps remain. First, substantially greater volumes of high-quality data are needed, particularly recordings with finer spatial and temporal resolution, extended temporal coverage, and integration across modalities and scales. Second, the field requires modeling frameworks that transcend linear mappings while retaining interpretability. Parameters and learned representations must remain open to mechanistic interrogation rather than functioning as opaque predictors. Third, progress in measurement and modeling must advance alongside hardware development. Next-generation sensing-enabled implants, high-density electrode arrays, and wearable systems capable of capturing behavior at sufficient resolution will all be essential to constrain richer models.

4.3 Looking Forward

These technical imperatives coincide with a broader transformation in scientific practice. The classical paradigm of hypothesis-driven experimentation, while foundational, explores only a vanishingly small region of the hypothesis space for systems as complex as the brain. Data-driven discovery—enabled by scalable measurement, expressive modeling, and rigorous held-out evaluation—offers a complementary epistemology, one capable of surfacing structure that no investigator would anticipate. This thesis sits at that interface: hypotheses guide what is measured and how models are constrained, but the models themselves surface patterns that suggest new questions.

This expanded reach carries commensurate responsibility. As neural interfaces grow more capable, questions of neural privacy, agency, and data governance will become increasingly urgent. The same tools that enable principled neuroscientific discovery could, if misapplied, facilitate surveillance of cognitive states or exploitation of neural data without meaningful consent. Our generation of scientists and engineers must build not only the technical infrastructure for understanding the brain, but also the ethical and institutional safeguards to ensure that such understanding serves human flourishing.

This thesis represents one set of steps along that path: showing that interpretable machine learning, grounded in naturalistic data and disciplined by generalization criteria, can advance both basic understanding of cortical representation and the translational

neuroscience of therapeutic neuromodulation. The work is incomplete; it raises more questions than it answers. The coming decades will demand that we measure more densely, model more thoughtfully, and reason more expansively about what it means to “understand” the brain with the responsibility that such understanding demands.

References

- Albin, R. L., Young, A. B., and Penney, J. B. (1989). The functional anatomy of basal ganglia disorders. *Trends in Neurosciences*, 12(10):366–375.
- Ansó, J., Benjaber, M., Parks, B., Parker, S., Oehr, C. R., Petrucci, M., Gilron, R., Little, S., Wilt, R., Bronte-Stewart, H., Gunduz, A., Borton, D., Starr, P. A., and Denison, T. (2022). Concurrent stimulation and sensing in bi-directional brain interfaces: a multi-site translational experience. *Journal of Neural Engineering*, 19(2):026025.
- Antonello, R., Vaidya, A., and Huth, A. (2023). Scaling laws for language encoding models in fMRI. *Advances in Neural Information Processing Systems*, 36:21895–21907.
- Arora, S., Li, Y., Liang, Y., Ma, T., and Risteski, A. (2016). A latent variable model approach to PMI-based word embeddings. *Transactions of the Association for Computational Linguistics*, 4:385–399.
- Arora, S., Li, Y., Liang, Y., Ma, T., and Risteski, A. (2018). Linear algebraic structure of word senses, with applications to polysemy.
- Ashkan, K., Rogers, P., Bergman, H., and Ughratdar, I. (2017). Insights into the mechanisms of deep brain stimulation. *Nature Reviews Neurology*, 13(9):548–554.
- Balogh, V., Berend, G., Diochnos, D. I., and Turán, G. (2020). Understanding the semantic content of sparse word embeddings using a commonsense knowledge base. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(5):7399–7406.
- Bau, D., Zhou, B., Khosla, A., Oliva, A., and Torralba, A. (2017). Network dissection: Quantifying interpretability of deep visual representations.
- Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1):289–300.

- Bergman, H., Wichmann, T., and DeLong, M. R. (1990). Reversal of experimental parkinsonism by lesions of the subthalamic nucleus. *Science*, 249(4975):1436–1438.
- Beuter, A. and Vasilakos, K. (1995). Tremor: is parkinson’s disease a dynamical disease? *Chaos: An interdisciplinary journal of nonlinear science*, 5(1):35–42.
- Bonnasse-Gahot, L. and Pallier, C. (2024). fmri predictors based on language models of increasing complexity recover brain left lateralization. *Advances in Neural Information Processing Systems*, 37:125231–125263.
- Bricken, T., Templeton, A., Batson, J., Chen, B., Jermyn, A., Conerly, T., Turner, N., Anil, C., Denison, C., Askell, A., Lasenby, R., Wu, Y., Kravec, S., Schiefer, N., Maxwell, T., Joseph, N., Hatfield-Dodds, Z., Tamkin, A., Nguyen, K., McLean, B., Burke, J. E., Hume, T., Carter, S., Henighan, T., and Olah, C. (2023). Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*.
- Brittain, J.-S. and Brown, P. (2014). Oscillations and the basal ganglia: motor control and beyond. *NeuroImage*, 85 Pt 2:637–647.
- Brown, P. (2003). Oscillatory nature of human basal ganglia activity: Relationship to the pathophysiology of parkinson’s disease. *Movement Disorders*, 18(4):357–363.
- Caligiore, D., Helmich, R. C., Hallett, M., Moustafa, A. A., Timmermann, L., Toni, I., and Baldassarre, G. (2016). Parkinson’s disease as a system-level disorder. *NPJ Parkinson’s disease*, 2:16025.
- Cona, G., Wiener, M., et al. (2024). A gradient of overlapping representations of time and space in the human cortex. *Journal of Neuroscience*.
- Connolly, B. S. and Lang, A. E. (2014). Pharmacological treatment of parkinson disease: a review. *Jama*, 311(16):1670–1683.
- Cunningham, H., Ewart, A., Riggs, L., Huben, R., and Sharkey, L. (2023). Sparse autoencoders find highly interpretable features in language models.
- Dehaene, S. (1997). *The number sense: how the mind creates mathematics*. Oxford University Press, New York.
- Dehaene, S. and Brannon, E. M. (2010). Space, time, and number: a kantian research program. *Trends in Cognitive Sciences*, 14(12):517–519.
- DeLong, M. R. (1990). Primate models of movement disorders of basal ganglia origin. *Trends in neurosciences*, 13(7):281–285.

- Deng, D., Ostrem, J. L., Nguyen, V., Cummins, D. D., Sun, J., Pathak, A., Little, S., and Abbasi-Asl, R. (2024). Interpretable video-based tracking and quantification of parkinsonism clinical motor states. *npj Parkinson's Disease*, 10(1):122.
- Deniz, F., Nunez-Elizalde, A. O., Huth, A. G., and Gallant, J. L. (2019). The representation of semantic information across human cerebral cortex during listening versus reading is invariant to stimulus modality. *Journal of Neuroscience*, 39(39):7722–7736.
- Dixon, T. C., Strandquist, G., Zeng, A., Frączek, T., Bechtold, R., Lawrence, D., Ravi, S., Starr, P. A., Gallant, J. L., Herron, J. A., et al. (2025). Movement-responsive deep brain stimulation for parkinson's disease using a remotely optimized neural decoder. *Nature Biomedical Engineering*, pages 1–15.
- Dupré la Tour, T., Eickenberg, M., Nunez-Elizalde, A. O., and Gallant, J. L. (2022). Feature-space selection with banded ridge regression. *NeuroImage*, 264:119728.
- Dupré la Tour, T., Visconti di Oleggio Castello, M., and Gallant, J. L. (2025). The voxelwise encoding model framework: A tutorial introduction to fitting encoding models to fMRI data. *Imaging Neuroscience*, 3:imag_a_00575.
- Elhage, N., Hume, T., Olsson, C., Schiefer, N., Henighan, T., Kravec, S., Hatfield-Dodds, Z., Lasenby, R., Drain, D., Chen, C., Grosse, R., McCandlish, S., Kaplan, J., Amodei, D., Wattenberg, M., and Olah, C. (2022). Toy models of superposition. *Transformer Circuits Thread*.
- Faruqui, M., Tsvetkov, Y., Yogatama, D., Dyer, C., and Smith, N. A. (2015). Sparse over-complete word vector representations. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics*, pages 1491–1500, Beijing, China. Association for Computational Linguistics.
- Gao, J. S., Huth, A. G., Lescroart, M. D., and Gallant, J. L. (2015). Pycortex: an interactive surface visualizer for fMRI. *Frontiers in Neuroinformatics*, 9.
- Gilron, R., Little, S., Perrone, R., Wilt, R., De Hemptinne, C., Yaroshinsky, M. S., Racine, C. A., Wang, S. S., Ostrem, J. L., Larson, P. S., Wang, D. D., Galifianakis, N. B., Bledsoe, I. O., San Luciano, M., Dawes, H. E., Worrell, G. A., Kremen, V., Borton, D. A., Denison, T., and Starr, P. A. (2021). Long-term wireless streaming of neural recordings for circuit discovery and adaptive stimulation in individuals with parkinson's disease. *Nature Biotechnology*, 39(9):1078–1085.
- Goering, S. et al. (2021). Recommendations for responsible development and application of neurotechnologies. *Neuroethics*.

- Goetz, C. G., Tilley, B. C., Shaftman, S. R., et al. (2008). Movement Disorder Society-sponsored revision of the Unified Parkinson's Disease Rating Scale (MDS-UPDRS): Scale presentation and clinimetric testing results. *Movement Disorders*, 23(15):2129–2170.
- Goldstein, A., Zada, Z., Buchnik, E., Schain, M., Price, A., et al. (2022). Shared computational principles for language processing in humans and deep language models. *Nature Neuroscience*, 25(3):369–380.
- Gorzelic, P., Schiff, S. J., and Sinha, A. (2013). Model-based rational feedback controller design for closed-loop deep brain stimulation of parkinson's disease. *Journal of Neural Engineering*, 10(2):026016.
- Grill, W. M., Snyder, A. N., and Miocinovic, S. (2004). Deep brain stimulation creates an informational lesion of the stimulated nucleus. *Neuroreport*, 15(7):1137–1140.
- Gurnee, W., Nanda, N., Pauly, M., Harvey, K., Troitskii, D., and Bertsimas, D. (2023). Finding neurons in a haystack: Case studies with sparse probing.
- Hashimoto, T., Elder, C. M., Okun, M. S., Patrick, S. K., and Vitek, J. L. (2003). Stimulation of the subthalamic nucleus changes the firing pattern of pallidal neurons. *Journal of Neuroscience*, 23(5):1916–1923.
- Huth, A. G., de Heer, W. A., Griffiths, T. L., Theunissen, F. E., and Gallant, J. L. (2016). Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature*, 532(7600):453–458.
- Huth, A. G., Nishimoto, S., Vu, A. T., and Gallant, J. L. (2012). A continuous semantic space describes the representation of thousands of object and action categories across the human brain. *Neuron*, 76(6):1210–1224.
- Jain, S. and Huth, A. (2018). Incorporating context into language encoding models for fMRI. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.
- Karpathy, A., Johnson, J., and Fei-Fei, L. (2015). Visualizing and understanding recurrent networks.
- Khanna, P. and Carmena, J. M. (2015). Neural oscillations: beta band activity across motor networks. *Current Opinion in Neurobiology*, 32:60–67.
- Khanna, P. and Carmena, J. M. (2017). Beta band oscillations in motor cortex reflect neural population signals that delay movement onset. *eLife*, 6:e24573.

- Koeglsperger, T., Palleis, C., Hell, F., Mehrkens, J. H., and Bötzel, K. (2019). Deep brain stimulation programming for movement disorders: Current concepts and evidence-based strategies. *Frontiers in Neurology*, 10.
- Kumar, S., Sumers, T. R., Yamakoshi, T., Goldstein, A., Hasson, U., Norman, K. A., Griffiths, T. L., Hawkins, R. D., and Nastase, S. A. (2023). Shared functional specialization in transformer-based language models and the human brain.
- La Tour, T. D., Eickenberg, M., Nunez-Elizalde, A. O., and Gallant, J. L. (2022). Feature-space selection with banded ridge regression. *NeuroImage*, 264:119728.
- la Tour, T. D., Lu, M., Eickenberg, M., and Gallant, J. L. (2021). A finer mapping of convolutional neural network layers to the visual cortex. In *SVRHM 2021 Workshop @ NeurIPS*.
- Leonard, M. K., Gwilliams, L., Sellers, K. K., Chung, J. E., et al. (2024). Large-scale single-neuron speech sound encoding across the depth of human cortex. *Nature*, 626(7999):593–602.
- Leventhal, D. K., Gage, G. J., Schmidt, R., Pettibone, J. R., Case, A. C., and Berke, J. D. (2012). Basal ganglia beta oscillations accompany cue utilization. *Neuron*, 73(3):523–536.
- Li, Y., Anumanchipalli, G. K., Mohamed, A., Chen, P., Carney, L. H., Lu, J., Wu, J., and Chang, E. F. (2023). Dissecting neural computations in the human auditory pathway using deep neural networks for speech. *Nature Neuroscience*, 26(12):2213–2225.
- Little, S., Beudel, M., Zrinzo, L., Foltynie, T., Limousin, P., Hariz, M., Neal, S., Cheeran, B., Cagnan, H., Gratwicke, J., Aziz, T. Z., Pogosyan, A., and Brown, P. (2016). Bilateral adaptive deep brain stimulation is effective in parkinson’s disease. *Journal of Neurology, Neurosurgery & Psychiatry*, 87(7):717–721.
- Little, S. and Brown, P. (2012). What brain signals are suitable for feedback control of deep brain stimulation in parkinson’s disease? *Annals of the New York Academy of Sciences*, 1265(1):9–24.
- Little, S. and Brown, P. (2014). The functional role of beta oscillations in parkinson’s disease. *Parkinsonism and Related Disorders*, 20:S44–S48.
- Little, S., Pogosyan, A., Kuhn, A. A., and Brown, P. (2012). Beta band stability over time correlates with parkinsonian rigidity and bradykinesia. *Experimental Neurology*, 236(2):383–388.

- Littlejohn, K. T., Cho, C. J., Liu, J. R., Silva, A. B., et al. (2025). A streaming brain-to-voice neuroprosthesis to restore naturalistic communication. *Nature Neuroscience*, 28(4):902–912.
- Lozano, A. M., Lipsman, N., Bergman, H., Brown, P., Chabardes, S., Chang, J. W., Matthews, K., McIntyre, C. C., Schlaepfer, T. E., Schulder, M., Temel, Y., Volkmann, J., and Krauss, J. K. (2019). Deep brain stimulation: current challenges and future directions. *Nature Reviews Neurology*, 15(3):148–160.
- Lugaresi, C., Tang, J., Nash, H., McClanahan, C., Uboweja, E., Hays, M., Zhang, F., Chang, C.-L., Yong, M. G., Lee, J., et al. (2019). Mediapipe: A framework for building perception pipelines. *arXiv preprint arXiv:1906.08172*.
- Mathiopoulou, V., Habets, J., Feldmann, L. K., Busch, J. L., Roediger, J., Behnke, J. K., Schneider, G.-H., Faust, K., and Kühn, A. A. (2025). Gamma entrainment induced by deep brain stimulation as a biomarker for motor improvement with neuromodulation. *Nature Communications*, 16(1):2956.
- Mathis, M. W., Rotondo, A. P., Chang, E. F., Tolias, A. S., and Mathis, A. (2024). Decoding the brain: From neural representations to mechanistic models. *Cell*, 187(21):5814–5832.
- McGregor, M. M. and Nelson, A. B. (2019). Circuit mechanisms of parkinson’s disease. *Neuron*, 101(6):1042–1056.
- Meschke, E. X., Castello, M. V. d. O., Tour, T. D. I., and Gallant, J. L. (2023). Model connectivity: leveraging the power of encoding models to overcome the limitations of functional connectivity.
- Mikolov, T., Yih, W.-t., and Zweig, G. (2013). Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 746–751, Atlanta, Georgia. Association for Computational Linguistics.
- Miller, G. A. (1995). WordNet: a lexical database for english. *Commun. ACM*, 38(11):39–41.
- Nakai, T. and Nishimoto, S. (2020). Quantitative models of visual brain representations. *NeuroImage*.
- Neumann, W.-J., Steiner, L. A., and Milosevic, L. (2023). Neurophysiological mechanisms of deep brain stimulation across spatiotemporal resolutions. *Brain*, 146(11):4456–4468.
- Nunez-Elizalde, A. O., Huth, A. G., and Gallant, J. L. (2019). Voxelwise encoding models with non-spherical multivariate normal priors. *NeuroImage*, 197:482–492.

- Oehrn, C. R., Cernera, S., Hammer, L. H., Shcherbakova, M., Yao, J., Hahn, A., Wang, S., Ostrem, J. L., Little, S., and Starr, P. A. (2024). Chronic adaptive deep brain stimulation versus conventional stimulation in parkinson's disease: a blinded randomized feasibility trial. *Nature Medicine*, 30(11):3345–3356.
- Olah, C., Mordvintsev, A., and Schubert, L. (2017). Feature visualization. *Distill*, 2(11):10.23915/distill.00007.
- Olaru, M., Cernera, S., Hahn, A., Wozny, T. A., Anso, J., de Hemptinne, C., Little, S., Neumann, W.-J., Abbasi-Asl, R., and Starr, P. A. (2024). Motor network gamma oscillations in chronic home recordings predict dyskinesia in parkinson's disease. *Brain*, 147(6):2038–2052.
- Olaru, M., Hahn, A., Shcherbakova, M., Little, S., Neumann, W.-J., Abbasi-Asl, R., and Starr, P. A. (2025). Deep brain stimulation-entrained gamma oscillations in chronic home recordings in parkinson's disease. *Brain Stimulation*, 18(2):132–141.
- Oliver, M., Winter, M., Tour, T. D. I., Eickenberg, M., and Gallant, J. L. (2024). A biologically-inspired hierarchical convolutional energy model predicts v4 responses to natural videos.
- Olshausen, B. A. and Field, D. J. (1996). Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381(6583):607–609.
- Olshausen, B. A. and Field, D. J. (1997). Sparse coding with an overcomplete basis set: A strategy employed by v1? *Vision Research*, 37(23):3311–3325.
- Ou, Z., Pan, J., Tang, S., Duan, D., Yu, D., Nong, H., and Wang, Z. (2021). Global trends in the incidence, prevalence, and years lived with disability of parkinson's disease in 204 countries/territories from 1990 to 2019. *Frontiers in public health*, 9:776847.
- Pedersen, T., Patwardhan, S., and Michelizzi, J. (2004). WordNet::similarity - measuring the relatedness of concepts. In *Demonstration Papers at HLT-NAACL 2004*, pages 38–41, Boston, Massachusetts, USA. Association for Computational Linguistics.
- Pennington, J., Socher, R., and Manning, C. (2014). GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.
- Phipson, B. and Smyth, G. K. (2010). Permutation p-values should never be zero: calculating exact p-values when permutations are randomly drawn. *Statistical Applications in Genetics and Molecular Biology*, 9:Article39.

- Poewe, W., Seppi, K., Tanner, C. M., Halliday, G. M., Brundin, P., Volkmann, J., Schrag, A.-E., and Lang, A. E. (2017). Parkinson disease. *Nature Reviews Disease Primers*, 3(1):1–21.
- Pogosyan, A., Yoshida, F., Chen, C. C., Martinez-Torres, I., Foltynie, T., Limousin, P., Zrinzo, L., Hariz, M. I., and Brown, P. (2010). Parkinsonian impairment correlates with spatially extensive subthalamic oscillatory synchronization. *Neuroscience*, 171(1):245–257.
- Rubin, J. E. and Terman, D. (2004). High frequency stimulation of the subthalamic nucleus eliminates pathological thalamic rhythmicity in a computational model. *Journal of Computational Neuroscience*, 16(3):211–235.
- Räuker, T., Ho, A., Casper, S., and Hadfield-Menell, D. (2023). Toward transparent AI: A survey on interpreting the inner structures of deep neural networks.
- Schakel, A. M. J. and Wilson, B. J. (2015). Measuring word significance using distributed representations of words.
- Schiff, S. J. (2010). Towards model-based control of parkinson’s disease. *Philosophical transactions. Series A, Mathematical, physical, and engineering sciences*, 368(1918):2269–2308.
- Schrimpf, M., Blank, I., Tuckute, G., Kauf, C., Hosseini, E. A., Kanwisher, N., Tenenbaum, J., and Fedorenko, E. (2020). Artificial neural networks accurately predict language processing in the brain.
- Sellers, K. K., Gilron, R., Anso, J., Louie, K. H., Shirvalkar, P. R., Chang, E. F., Little, S. J., and Starr, P. A. (2021). Analysis-rcs-data: open-source toolbox for the ingestion, time-alignment, and visualization of sense and stimulation data from the medtronic summit rc+ s system. *Frontiers in Human Neuroscience*, 15:714256.
- Sermon, J. J., Olaru, M., Ansó, J., Cernera, S., Little, S., Shcherbakova, M., Bogacz, R., Starr, P. A., Denison, T., and Duchet, B. (2023). Sub-harmonic entrainment of cortical gamma oscillations to deep brain stimulation in parkinson’s disease: Model based predictions and validation in three human subjects. *Brain Stimulation*, 16(5):1412–1424.
- Sharkey, L., Chughtai, B., Batson, J., Lindsey, J., Wu, J., Bushnaq, L., Goldowsky-Dill, N., Heimersheim, S., Ortega, A., Bloom, J., Biderman, S., Garriga-Alonso, A., Conmy, A., Nanda, N., Rumbelow, J., Wattenberg, M., Schoots, N., Miller, J., Michaud, E. J., Casper, S., Tegmark, M., Saunders, W., Bau, D., Todd, E., Geiger, A., Geva, M., Hoogland, J., Murfet, D., and McGrath, T. (2025). Open problems in mechanistic interpretability.
- Shenoy, K. V., Sahani, M., and Churchland, M. M. (2013). Cortical control of arm movements: A dynamical systems perspective. *Annual Review of Neuroscience*, 36(1):337–359.

- Silva, A. B., Liu, J. R., Metzger, S. L., Bhaya-Grossman, I., et al. (2024). A bilingual speech neuroprosthesis driven by cortical articulatory representations shared between languages. *Nature Biomedical Engineering*, 8(8):977–991.
- Simsek, O. et al. (2025). Neural overlaps of time, space, and number. *Cognitive Neuropsychology*.
- Skagerlund, K., Karlsson, T., and Träff, U. (2016). Magnitude processing in the brain: An fMRI study of time, space, and numerosity as a shared cortical system. *Frontiers in Human Neuroscience*, 10.
- Spooner, R. K., Hizli, B. J., Bahners, B. H., Schnitzler, A., and Florin, E. (2024). Modulation of DBS-induced cortical responses and movement by the directionality and magnitude of current administered. *npj Parkinson's Disease*, 10(1):53.
- Stanslaski, S., Herron, J., Chouinard, T., Bourget, D., Isaacson, B., Kremen, V., Opri, E., Drew, W., Brinkmann, B. H., Gunduz, A., Adamski, T., Worrell, G. A., and Denison, T. (2018). A chronically implantable neural coprocessor for investigating the treatment of neurological disorders. *IEEE transactions on biomedical circuits and systems*, 12(6):1230–1245.
- Starr, P. A. (2018). Totally implantable bidirectional neural prostheses: A flexible platform for innovation in neuromodulation. *Frontiers in Neuroscience*, 12:619.
- Stein, E. and Bar-Gad, I. (2013). Beta oscillations in the cortico-basal ganglia loop during parkinsonism. *Experimental Neurology*, 245:52–59.
- Strandquist, G., Frączek, T., Dixon, T., Ravi, S., Bechtold, R., Lawrence, D., Zeng, A., Gallant, J., Little, S., and Herron, J. (2023). Bringing the clinic home: An at-home multi-modal data collection ecosystem to support adaptive deep brain stimulation. *Journal of Visualized Experiments: JoVE*.
- Tang, J., Du, M., Vo, V., Lal, V., and Huth, A. (2023a). Brain encoding models based on multimodal transformers can transfer across language and vision. *Advances in Neural Information Processing Systems*, 36:29654–29666.
- Tang, J., LeBel, A., Jain, S., and Huth, A. G. (2023b). Semantic reconstruction of continuous language from non-invasive brain recordings. *Nature Neuroscience*, 26(5):858–866.
- Thanvi, B., Lo, N., and Robinson, T. (2007). Levodopa-induced dyskinesia in parkinson's disease: clinical features, pathogenesis, prevention and treatment. *Postgraduate medical journal*, 83(980):384–388.

- Walsh, V. (2003). A theory of magnitude: common cortical metrics of time, space and quantity. *Trends in Cognitive Sciences*, 7(11):483–488.
- Wehbe, L., Murphy, B., Talukdar, P., Fyshe, A., Ramdas, A., and Mitchell, T. (2014a). Simultaneously uncovering the patterns of brain regions involved in different story reading subprocesses. *PloS One*, 9(11):e112575.
- Wehbe, L., Vaswani, A., Knight, K., and Mitchell, T. (2014b). Aligning context-based statistical models of language with brain activity during reading. In Moschitti, A., Pang, B., and Daelemans, W., editors, *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 233–243, Doha, Qatar. Association for Computational Linguistics.
- Williams, R. (2010). Alim-louis benabid: stimulation and serendipity. *The Lancet Neurology*, 9(12):1152.
- Wu, M. C.-K., David, S. V., and Gallant, J. L. (2006). Complete functional characterization of sensory neurons by system identification. *Annual Review of Neuroscience*, 29:477–505.
- Wu, Z. and Palmer, M. (1994). Verb semantics and lexical selection. In *32nd Annual Meeting of the Association for Computational Linguistics*, pages 133–138, Las Cruces, New Mexico, USA. Association for Computational Linguistics.
- Yamins, D. L. K. and DiCarlo, J. J. (2016). Using goal-driven deep learning models to understand sensory cortex. *Nature Neuroscience*, 19(3):356–365.
- Yu, Y., Sanabria, D. E., Wang, J., Hendrix, C. M., Zhang, J., Nebeck, S. D., Amundson, A. M., Busby, Z. B., Bauer, D. L., Johnson, M. D., Johnson, L. A., and Vitek, J. L. (2021). Parkinsonism alters beta burst dynamics across the basal ganglia–motor cortical network. *Journal of Neuroscience*, 41(10):2274–2286.
- Yun, Z., Chen, Y., Olshausen, B. A., and LeCun, Y. (2023). Transformer visualization via dictionary learning: contextualized embedding as a linear superposition of transformer factors.
- Zeng, A. and Gallant, J. L. (2025). Disentangling superpositions: Interpretable brain encoding model with sparse concept atoms. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.
- Zhang, J., Chen, Y., Cheung, B., and Olshausen, B. A. (2021). Word embedding visualization via dictionary learning.