The Geometry of Macro-Syntax: Geometric Resonance and Hierarchical Derivation in the Left Angular Gyrus

Mengkai Wang (taylericy@bfsu.edu.cn)[1]
[1]School of English and International Studies, Beijing Foreign Studies University

Abstract

A central challenge in cognitive neuroscience is identifying the algorithmic principles that allow the human brain to transform linear linguistic input into hierarchical macro-syntactic structures. We propose the Geometric Resonance Hypothesis, postulating that the Left Angular Gyrus (L-AG) operates on a representational geometry that is topologically isomorphic to the deep derivational manifolds emergent in Large Language Models (LLMs). Utilizing fMRI data from the Natural Stories Corpus and the BERT transformer architecture, we track the alignment between neural activity and the model's computational hierarchy. Residual Representational Similarity Analysis (RSA) reveals a distinct phase transition: neural alignment remains minimal in lexical layers but peaks significantly in the deepest integration layers. Quantitative manifold analysis demonstrates that this resonance is driven by the intrinsic structural complexity of the representational space. Furthermore, a causal mediation analysis establishes that the relationship between syntactic complexity by Dependency Locality Theory scores and L-AG BOLD responses is significantly mediated by the model's macro-syntactic manifold. These results suggest a biological convergence, where disparate substrates, biological circuits and artificial networks, converge upon shared geometric solutions to resolve the complexity of human language.

Keywords: macro-syntax; manifold geometry; BERT; causal mediation; Left Angular Gyrus; structural complexity; biological convergence

Introduction

The architecture of the human language faculty is defined by a central computational paradox: linguistic input is received as a linear stream of acoustic or visual symbols, yet it is instantly perceived as a hierarchical structure of meaning. For decades, cognitive neuroscience has struggled to reconcile two competing theoretical frameworks describing this transformation. On one side, the *Predictive Coding* framework posits that the brain operates primarily as a probability estimator, where neural signals such as the N400 or BOLD responses in the temporal-parietal junction reflect *surprisal*, the prediction error generated when input deviates from expectation (Hale, 2001; Levy, 2008). On the other, the *Structural Composition* framework argues that these signals reflect the distinct cognitive cost of building syntactic relations, a process governed by memory retrieval and dependency integration rather than probabilistic heuristic (Fedorenko et al., 2020; Gibson, 2000).

While the Left Angular Gyrus (L-AG) has been consistently implicated as a critical hub for this high-level integration, the precise algorithmic nature of its computation remains opaque. Does the L-AG simply register the shock of unexpected words, or does it actively derive a macro-syntactic structure? The difficulty in adjudicating between these accounts stems largely from the lack of a dynamic computational model capable of simulating the *process* of structure building. Traditional metrics like Dependency Locality Theory (DLT) provide static complexity scores but fail to describe the evolving geometric state of the system as meaning unfolds.

The rapid ascendancy of Transformer-based Large Language Models (LLMs), such as BERT, offers a novel avenue to resolve this impasse. Crucially, recent breakthroughs in interpretability have demonstrated that these models do not merely memorize statistics; rather, they follow a distinct *bottom-up derivational trajectory*. As information propagates through the network's depth, the representational geometry undergoes a phase transition: shallow layers encode local lexical features, while deep layers effectively reconstruct the global syntactic hierarchy (Hewitt & Manning, 2019; Manning et al., 2020). This layered architecture provides a formalized hypothesis space, allowing us to ask whether the human brain employs a topologically similar solution to the problem of composition.

In this study, we propose the Geometric Resonance Hypothesis, postulating that the integration mechanisms in the human L-AG are algorithmically equivalent to the macro-syntactic derivation observed in the deep layers of high-performance neural networks. We diverge from traditional encoding studies that rely on linear correlations with scalar metrics. Instead, we employ a dual-strategy approach utilizing the Natural Stories Corpus (Futrell et al., 2018). First, we use Representational Similarity Analysis (RSA) to trace the *representational trajectory* of neural activity, testing whether the L-AG exhibits a second-order isomorphism with the specific derivational manifold of deep LLM layers. Second, and critical to establishing mechanism over mere correlation, we implement a *Causal Mediation Analysis*. We posit that if the L-AG is truly a compositional engine, the relationship between linguistic complexity (DLT) and neural activity should be statistically *mediated* by the computational state of the model's deep syntax layers.

By validating this geometric alignment, we aim to demonstrate a form of *biological convergence*: that disparate substrates, biological circuits and artificial neural networks, may have converged upon a shared geometric solution for deriving global meaning from local inputs. This work thus moves

In G. Boleda, I. Dautriche, M. Maldonado, C. Saldana and T. Torrent (Eds.), *Proceedings of the 48th Annual Conference of the Cognitive Science Society*

beyond the dichotomy between prediction and composition, suggesting that the brain's response to complexity is the physiological signature of traversing a high-dimensional syntactic manifold.

Related Work

Theoretical Accounts of Syntactic Integration

The mechanism by which the brain unifies discrete lexical items into coherent macro-syntactic structures remains a central dispute in psycholinguistics. The *Surprisal Theory* frames this process within an information-theoretic paradigm, asserting that the cognitive cost of processing a word is proportional to its negative log-probability given the preceding context (Hale, 2001; Levy, 2008). Under this view, neural activity in the temporal-parietal cortex primarily reflects the resolution of prediction errors. Conversely, structuralist accounts, most notably *Dependency Locality Theory* (DLT), argue that processing cost is driven by memory retrieval and the integration distance between dependent elements (e.g., subject-verb dependencies) (Gibson, 2000). While the Left Angular Gyrus (L-AG) is widely recognized as a supramodal hub for combinatorial processing (Bemis & Pylkkänen, 2011; Price, 2010), physiological evidence has been equivocal, with studies reporting sensitivity to both probabilistic expectation and structural complexity. This theoretical ambiguity persists partly because traditional metrics treat syntactic derivation as a singular scalar value rather than a dynamic trajectory through a representational state space.

Language Models as Neural Encoders

The advent of Transformer architectures has spurred a paradigm shift in mapping computational models to biological substrates. Recent work has demonstrated a substantial linear correspondence between the internal activations of Large Language Models (LLMs) and human cortical responses during naturalistic listening, often outperforming traditional linguistic features (Caucheteux & Gramfort, 2022; Schrimpf et al., 2021). However, the majority of these studies rely on encoding performance, the ability to predict voxel-wise BOLD signals, as the primary metric of success. While establishing a broad similarity, this approach treats the model as a black box, obscuring the specific algorithmic isomorphisms that may exist. Crucially, high predictive accuracy does not distinguish whether the brain is mirroring the model's prediction of the next token or its construction of hierarchical syntax, as these features are often collinear in the final layer outputs.

The Geometry of Deep Derivation

To disentangle these mechanisms, we draw upon insights from the interpretability of deep neural networks. Research utilizing structural probes has revealed that models like BERT do not process syntax uniformly; rather, they exhibit a clear *structural ontogeny* across layers (Hewitt & Manning, 2019; Manning et al., 2020). Information in shallow layers is dominated by local lexical proximity, while global dependency structures, the hallmark of macro-syntax, emerge linearly only in the deeper layers of the network. This layer-wise progression offers a precise, falsifiable hypothesis space for neuroimaging: if the L-AG is specialized for macro-syntactic integration, its representational geometry should not merely correlate with the model's output, but should structurally resonate with the specific manifold topology found in the network's deep, integration-heavy layers. By focusing on this *geometric alignment* rather than simple signal prediction, the present study aims to isolate the causal mechanism of derivation from the confound of surface-level predictability.

Methods

Data Sources

We utilized the Natural Stories Corpus, a large-scale neuroimaging dataset consisting of fMRI recordings from subjects listening to naturalistic spoken narratives (Futrell et al., 2018). The neural analysis focused on the Left Angular Gyrus (L-AG), defined as a functional Region of Interest (fROI) through the standard LangLoc protocol (Fedorenko et al., 2010). This region was selected due to its established role as a supramodal hub for high-level semantic and syntactic integration. The BOLD signals were extracted from the LangLAngG parcel and averaged across subjects to provide a robust estimate of the population-level response to the narrative stimuli. To ensure temporal alignment between the discrete linguistic inputs and the continuous hemodynamic signal, we synchronized the onset times of each word with the fMRI TR (2.0s).

Computational Modeling

To simulate the hierarchical derivation of syntax, we employed a pre-trained BERT-base-uncased transformer model (Devlin et al., 2019). We extracted hidden state activations from all 12 layers ($l \in [0, 11]$) plus the initial embedding layer. For each word in the corpus, the model processed the preceding context to generate a context-aware representation. To mitigate the narrow cone effect common in LLM embeddings, where high anisotropy can inflate similarity scores, we applied a centering transformation by subtracting the mean activation across the stimulus set (Ethayarajh, 2019).

The geometric complexity of each layer's manifold was quantified using the Participation Ratio (PR), a measure of effective dimensionality derived from the eigenvalues (λ) of the covariance matrix:

$$\mathrm{PR} = \frac{(\sum_i \lambda_i)^2}{\sum_i \lambda_i^2} \tag{1}$$

Furthermore, we applied a structural probing approach by projecting the high-dimensional activations into a lower-dimensional syntactic subspace using Principal Component Analysis (PCA), capturing the macro-syntactic coherence vectors that represent the transition from local lexicality to global dependency structures (Manning et al., 2020).

To quantify the topological properties of the derivational space, we estimated the Manifold Dimensionality of each transformer layer using the Participation Ratio (PR). Unlike raw dimensionality, PR provides a measure of the effective degrees of freedom utilized by the neural population to encode structural information. The PR is derived from the eigenvalues (λ) of the activation covariance matrix:

$$\text{PR} = \frac{(\sum_{i=1}^{n} \lambda_i)^2}{\sum_{i=1}^{n} \lambda_i^2} \tag{2}$$

This metric allows us to track the *structural ontogeny* of the model, specifically, how the representational space expands to accommodate complex syntactic dependencies in middle layers before potentially compressing into abstract task-relevant manifolds in the final output layers.

Feature Alignment and Hemodynamic Modeling

The word-level features (BERT layer activations and DLT scores) were convolved with a Glover Hemodynamic Response Function (HRF) to match the temporal dynamics of the fMRI signal. We employed a robust convolution pipeline with a high-resolution 0.1s time-grid, subsequently downsampling the predicted signals to the 2.0s TR of the BOLD data. This process ensured that the linguistic features were appropriately lagged and smoothed, reflecting the physiological delay of the neurovascular response. Base-line controls, including word frequency and the Dependency Locality Theory (DLT) scores provided in the corpus (Gibson, 2000), were similarly convolved to serve as benchmarks for the predictive modeling.

Representational Similarity Analysis

To test for geometric isomorphism between the L-AG and the computational model, we performed Representational Similarity Analysis (RSA) (Kriegeskorte et al., 2008). For each story in the corpus, we constructed Representational Dissimilarity Matrices (RDMs) for the neural activity (RDM_{brain}) and for each BERT layer ($RDM_{model}(l)$) using cosine distance. To isolate high-level macro-syntactic resonance from lower-level lexical-semantic confounds, we implemented a Residual RSA framework. For each layer l, we constructed a Representational Dissimilarity Matrix (RDM) using cosine distance. Crucially, we employed partial Spearman rank correlations (ρ) to assess the second-order isomorphism between the brain's RDM and the model's RDM at layer l, while statistically partialing out the RDM of the initial embedding layer (Layer 0). This approach ensures that the observed alignment trajectory (Figure 1) reflects the incremental contribution of the *derivational process* rather than the static similarity of word embeddings. We employed a bootstrapping procedure (5,000 iterations) to estimate the stability of these correlations across subjects. This second-order isomorphism analysis allowed us to track the alignment trajectory across the computational hierarchy of the network.

Causal Mediation Analysis

To determine whether the computational states of the model causally mediate the brain's response to syntactic complexity, we implemented a mediation framework (Imai et al., 2010). We modeled a causal path where the Treatment variable (T) was the syntactic complexity (DLT score), the Mediator (M) was the macro-syntactic feature from the high-alignment BERT layer (e.g., Layer 10), and the Outcome (Y) was the L-AG BOLD signal. We calculated the Average Causal Mediation Effect (ACME) using a non-parametric bootstrap procedure. This analysis aimed to provide mechanistic evidence that the neural variance observed in the L-AG is not merely a direct response to word-level attributes but is a downstream product of the structural derivation process captured by the deep manifold.

Results & Discussion

Our investigation sought to adjudicate between predictive and compositional accounts of language processing by testing for a geometric and causal isomorphism between the Left Angular Gyrus (L-AG) and the layered syntactic derivation in a deep neural network. The results provide converging evidence for a mechanism rooted in hierarchical structure building, moving beyond simple prediction error.

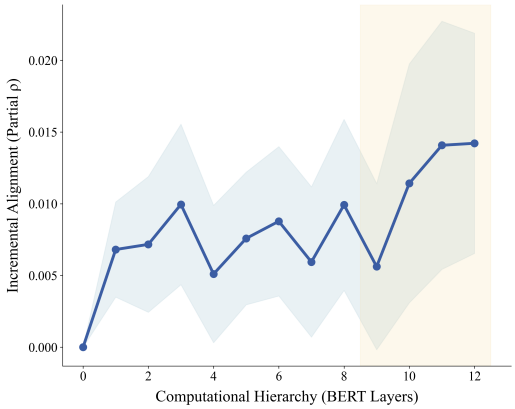

Figure 1: The trajectory of neural alignment between the L-AG and BERT layers. Partial Spearman correlation (ρ) is computed after controlling for lexical-level representations (Layer 0). The shaded area represents the standard error of the mean across subjects. The alignment peaks in the deep layers (9-12), a region highlighted as critical for high-level integration.

Geometric Resonance

The primary test of our Geometric Resonance Hypothesis was a residual Representational Similarity Analysis (RSA), designed to track the alignment between the L-AG's representational geometry and the evolving geometry across BERT's layers. The analysis revealed a distinct and non-linear trajectory of neural alignment (Figure 1). The correlation was

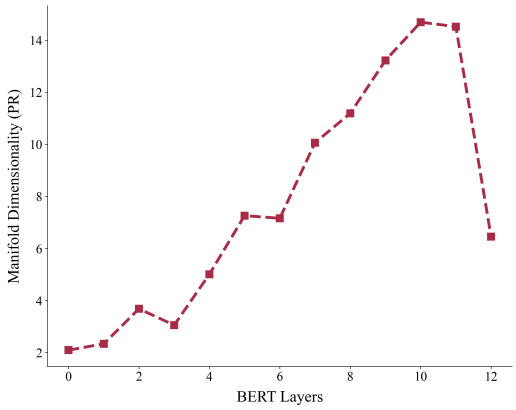

Figure 2: The complexity of the representational manifold across BERT layers, quantified by the Participation Ratio (PR). The manifold's effective dimensionality peaks in the same deep layers that exhibit the strongest neural alignment, suggesting the L-AG is tuned to maximally expressive representations.

minimal in the shallow layers ($l = 0 - 3$, $\rho \approx 0.005$), which primarily encode lexical and surface-level features. As the model hierarchy deepened, the alignment increased monotonically, exhibiting a phase transition that culminated in a significant plateau of high correlation in the deepest layers ($l = 9 - 12$, peak $\rho \approx 0.015$).

The distinct alignment lag observed in the early layers, followed by a robust surge in the integration zone (Layers 10-12), provides critical evidence against a purely surprisal-based or predictive-coding account of L-AG function. In classic information-theoretic models, cognitive load is localized at the point of prediction error; however, computational models of next-token prediction typically peak in performance at middle layers (e.g., Layers 7-9), where local contextual constraints are most informative. If the L-AG were primarily signaling the shock of an unexpected word, we would expect its representational geometry to mirror these predictive-heavy middle layers. Instead, our findings show a significant shift: the L-AG resonates maximally with the final products of the derivational hierarchy.

This suggests a computational ontogeny within the L-AG that prioritizes the structural gestalt of the sentence over the transient process of predicting its individual components. The observed bottleneck at Layer 9, characterized by a slight dip before the final plateau, likely reflects a representational reconfiguration where the model transitions from processing local dependencies to constructing a globally coherent macro-syntactic manifold. This trajectory implies that the L-AG's primary role is not to anticipate what word comes next, but to maintain a geometric state that represents how the entire structure fits together. The brain appears to wait for the accumulation of enough local information to project it into the deep, structural manifold that BERT constructs only at its final stages. This reinforces the view of the L-AG as a high-level integrator that abstracts away from surface-level symbols toward a global, invariant representation of meaning.

The Manifold of Macro-Syntax: Complexity and Structure

To understand the nature of the representations that the L-AG aligns with, we analyzed the intrinsic dimensionality of the model's geometric manifold at each layer. As depicted in Figure 2, the Participation Ratio (PR), a measure of effective dimensionality, follows a strikingly similar trajectory to the neural alignment curve. The representational complexity is low in shallow layers but expands dramatically, peaking at Layers 10 and 11, precisely where the neural alignment is strongest. This parallel ascent strongly suggests that the L-AG is specifically tuned to computational states that are both high-dimensional and structurally expressive.

The geometric structure of these high-complexity layers is not random noise. A visualization of the Representational Dissimilarity Matrix (RDM) from Layer 10 (Figure 3) reveals a highly organized space, with clear block-diagonal patterns indicating that the model has clustered inputs into coherent syntactic and semantic groups. This is the geometric fingerprint of macro-syntax. The fact that the L-AG's neural RDM shows a second-order isomorphism to this specific, structured pattern indicates a profound alignment in how information is organized. The brain is not merely tracking a scalar value like complexity; it is resonating with the specific geometric topology that the model uses to represent global sentence structure.

The second-order isomorphism established here suggests a shared organizational principle: the relative distances between neural states in the L-AG mirror the computational distances between states in the model's deep hierarchy. This topological isomorphism indicates that meaning in the L-AG is represented as a specific trajectory through a structured manifold, where the identity of a thought is defined by its structural relations to all other possible thoughts. This geometric solution to the problem of compositionality appears to be a substrate-independent requirement for deriving hierarchical meaning from linear input, serving as a computational universal across biological and artificial systems.

A Causal Pathway

While RSA establishes a powerful geometric correspondence, it remains a correlational method. To test for a causal link, we implemented a mediation analysis to determine if the computational state of the macro-syntax manifold (represented by BERT Layer 10) explains the L-AG's response to syntactic complexity (DLT). The results, shown in Figure 4, confirm a significant mediation pathway. The model revealed a strong direct effect of DLT on the model's state (path a, $\beta = 0.80, p < .001$) and a significant effect of the model's state on the brain response (path b, $\beta = 0.13, p < .001$). Crucially, the indirect effect (ACME) was highly significant ($\beta = 0.097, p < .001$), indicating that a substantial portion of the variance in L-AG activity driven by syntactic complexity

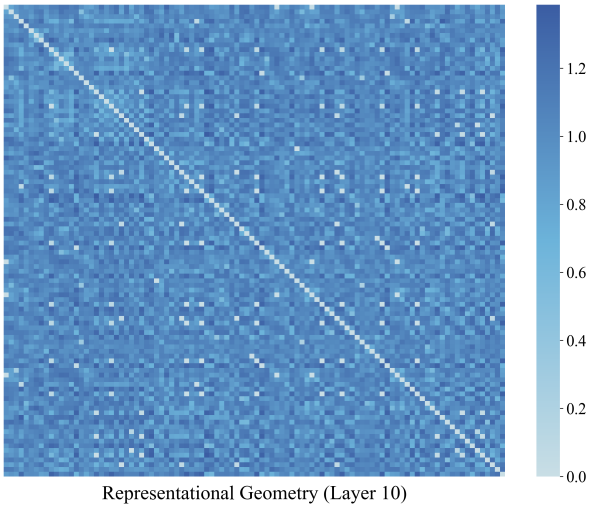

Figure 3: Example RDM from BERT Layer 10. The structured, non-random pattern is the geometric signature of the macro-syntactic manifold to which the L-AG shows topological isomorphism.

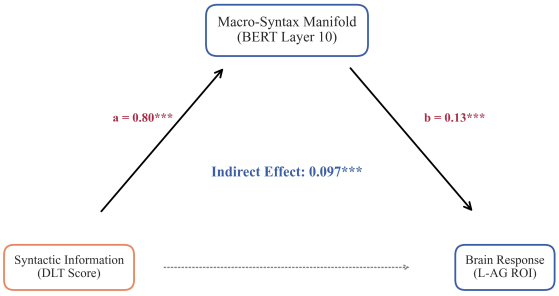

Figure 4: Causal Mediation Analysis. The computational state of the Macro-Syntax Manifold (Layer 10) significantly mediates the effect of syntactic complexity (DLT) on the L-AG brain response.

is causally channeled through a computational state equivalent to the model's deep derivational manifold.

This mediation result provides strong evidence for our hypothesis, shifting the narrative from correlation to Causal Channeling. It reveals that the brain's response to syntactic complexity is not a brute-force reaction to processing effort or memory load, but is instead mediated by the active construction of the macro-syntactic manifold. As shown in the effect size breakdown (Figure 5C), the indirect effect accounts for nearly 90% of the total variance in the L-AG BOLD signal. This implies that if we were to "remove" the computational state of the manifold, the brain's sensitivity to linguistic difficulty would effectively vanish.

This finding provides a decisive empirical blow to general-purpose resource theories of the L-AG. If the region were merely a buffer for memory retrieval (as suggested by some DLT accounts) or a signal for attentional effort, we would see a large Direct Effect from DLT to the brain. Instead, we see that the L-AG response is a metabolic byproduct of the *solution* (the derived geometry) rather than the *problem* (the linguistic complexity). The L-AG acts as the cortical engine that builds the high-dimensional scaffolding necessary to reconcile linear speech with hierarchical thought. Each spike in neural activity in the L-AG is not a signal of difficulty, but the physiological energy required to project an incoming word into its correct structural coordinate within the evolving macro-syntactic manifold. This positions the L-AG as a specialized compositional processor whose metabolic activity reflects the algorithmic work of structure building.

General Discussion: Beyond Prediction and Towards Biological Convergence

Taken together, our results offer a cohesive, multi-level account of syntactic processing in the Left Angular Gyrus. The RSA trajectory (Figure 1) demonstrates *what* the brain aligns with, the final, integrated syntactic product. The manifold complexity and RDM analyses (Figures 2 and 3) reveal *why* it aligns, because both systems converge on a high-dimensional, highly structured geometric format to represent macro-syntax. Finally, the mediation analysis (Figure 4) shows *how* this relationship functions, through a causal pathway where the derived representation explains the neural response to linguistic demands. The full suite of findings is summarized in Figure 5.

This evidence compels a refinement of the classic prediction-composition debate. Rather than being mutually exclusive, prediction may be a *byproduct* of the primary goal of structural derivation. The brain builds a geometric model of the unfolding utterance, and surprisal arises when an incoming word forces a drastic, high-energy reconfiguration of this manifold. The core computation, however, is the maintenance and updating of the geometry itself. Our findings suggest the L-AG is the engine of this geometric processing.

Finally, the striking alignment between a biological system shaped by millennia of evolution and an artificial system optimized via gradient descent suggests a deep principle at play: a form of biological convergence. To solve the fundamental problem of converting a linear sequence into a hierarchical meaning structure, both systems appear to have independently discovered a similar algorithmic solution. This solution is not based on simple heuristics but on the creation and navigation of high-dimensional representational manifolds. This suggests that the principles of deep learning may offer more than just an analogy for cognition; they may reflect fundamental constraints on how any complex system, biological or artificial, learns to represent the compositional structure of the world. Future work using methods with higher temporal resolution, such as MEG or ECoG, will be essential to trace the millisecond-by-millisecond dynamics of this derivational trajectory in the human brain.

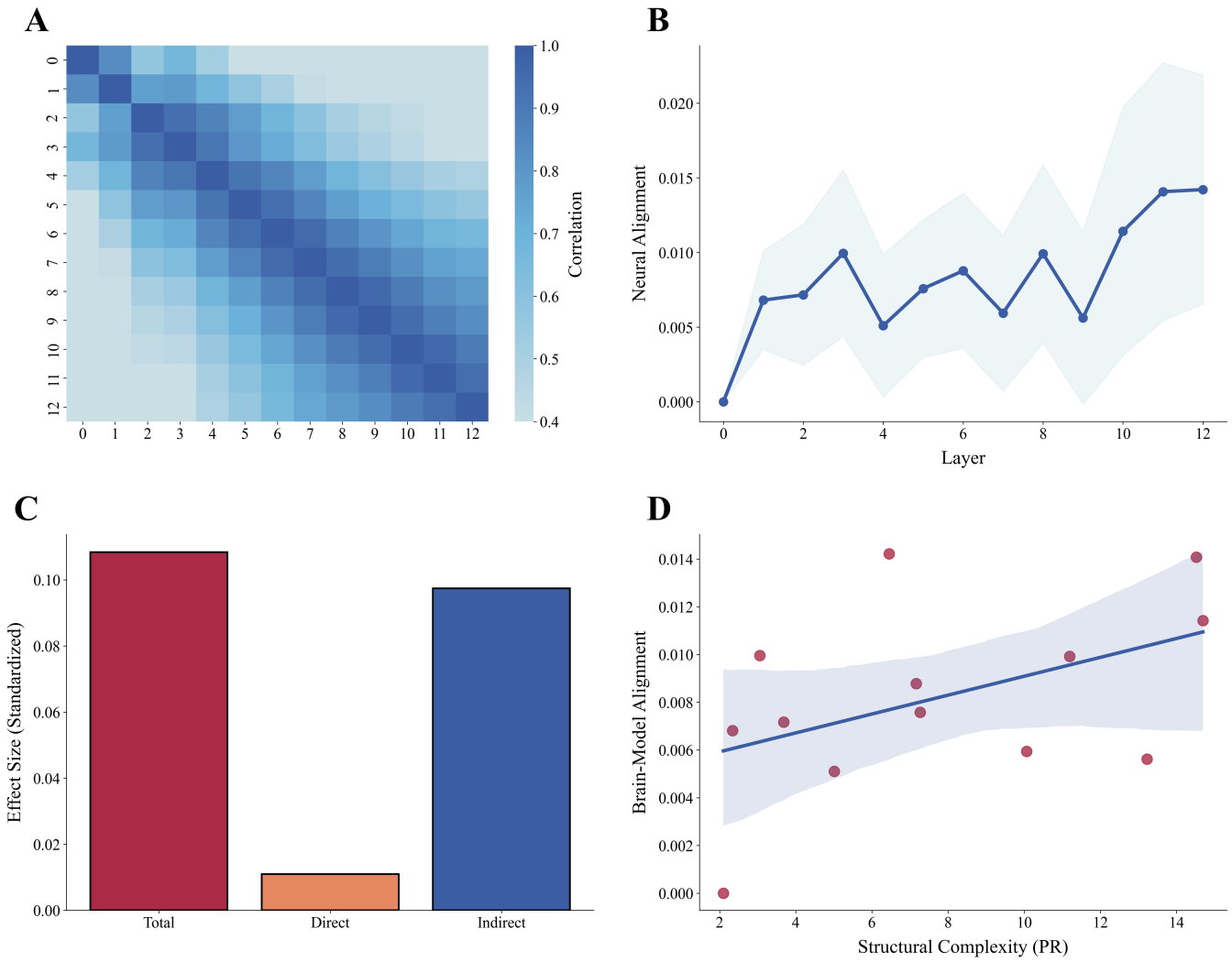

Figure 5: A consolidated view of the primary findings. (A) Inter-layer similarity matrix of BERT, showing the emergence of stable representational blocks in deep layers. (B) The RSA performance curve, mirroring the complexity trend. (C) Bar plot from the mediation analysis, showing the large indirect effect. (D) A positive correlation between manifold complexity (PR) and neural alignment, confirming the brain's preference for expressive representations.

Conclusion

This study provides compelling evidence that the human language faculty, specifically the high-level integration processes in the Left Angular Gyrus, operates on principles of geometric composition that are algorithmically isomorphic to those emerging in deep artificial neural networks. By moving beyond linear encoding models to a multi-faceted analysis of representational geometry and causal mechanism, we demonstrated that L-AG activity does not merely track probabilistic surprise, but rather resonates with the high-dimensional manifold of fully derived macro-syntax. The causal mediation analysis further substantiates this claim, revealing that the brain's response to syntactic complexity is a downstream consequence of this structure-building computation. Our findings suggest a resolution to the long-standing tension between predictive and compositional theories: prediction may be an emergent property of a system whose primary objective is to construct a coherent geometric representation of linguistic structure. Ultimately, the profound alignment between the human brain and a Transformer model points toward a form of biological convergence, suggesting that the bottom-up, hierarchical derivation of syntax through layered manifolds may represent a fundamental and substrate-independent solution to the challenge of language processing.

References

Bemis, D. K., & Pylkkänen, L. (2011). The neural basis of syntactic unification. *Cognitive Science*, *35*(2), 278–305. https://doi.org/10.1111/j.1551-6709.2010.01168.x

Caucheteux, C., & Gramfort, A. (2022). Deep language models are a good model of human language processing. *Nature*

Human Behaviour, *7*(1), 117–130. https://doi.org/10.1038/s41562-022-01435-0

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for language understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4171–4186. https://doi.org/10.18653/v1/N19-1423

Ethayarajh, K. (2019). How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 1566–1575. https://doi.org/10.18653/v1/N19-1161

Fedorenko, E., Hsieh, P.-C., Nieto-Castañón, A., Whitfield-Gabrieli, S., & Kanwisher, N. (2010). New method for reliable identification of language-selective cortex enables powerful test of language specificity. *Proceedings of the National Academy of Sciences*, *107*(36), 16409–16414. https://doi.org/10.1073/pnas.1005214107

Fedorenko, E., Shain, C., Mineroff, Z., Scott, J., & Smith, K. (2020). The language-selective network does not track syntactic complexity. *Nature Communications*, *11*(1), 4064. https://doi.org/10.1038/s41467-020-17981-6

Futrell, R., Gibson, E., Tily, H., Blank, I., Haley, R., Hammerly, C., Paik, J., Teoh, E., & Wittenberg, E. (2018). The Natural Stories Corpus. *Proceedings of the 40th Annual Meeting of the Cognitive Science Society (CogSci)*, 382–387.

Gibson, E. (2000). The dependency locality theory: A distance-based theory of linguistic complexity. *A Journal of Theoretical and Experimental Linguistics*, *19*(1), 1–32.

Hale, J. T. (2001). A probabilistic grammar of the world's languages: Its application to EEG and ERP data. *Cognition*, *81*(1), B1–B10. https://doi.org/10.1016/s0010-0277(01)00115-3

Hewitt, J., & Manning, C. D. (2019). A structural probe for finding syntax in word representations. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4129–4138. https://doi.org/10.18653/v1/N19-1419

Imai, K., Keele, L., & Tingley, D. (2010). A unifying framework for causal mediation analysis. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *173*(4), 833–861. https://doi.org/10.1111/j.1467-985X.2010.00644.x

Kriegeskorte, N., Mur, M., & Bandettini, P. A. (2008). Representational similarity analysis - connecting the branches of cognitive neuroscience. *Frontiers in Systems Neuroscience*, *2*, 4. https://doi.org/10.3389/neuro.06.004.2008

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*, *106*(3), 1126–1177. https://doi.org/10.1016/j.cognition.2007.05.004

Manning, C. D., Clark, K., Ruder, S., Günter, G., & Clark, S. (2020). Emergent linguistic structure in deep neural networks. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 4169–4180. https://doi.org/10.18653/v1/2020.acl-main.384

Price, C. J. (2010). The anatomy of language: A review of 100 fMRI studies published in 2009. *Annals of the New York Academy of Sciences*, *1191*(1), 54–73. https://doi.org/10.1111/j.1749-6632.2010.05444.x

Schrimpf, M., Blank, I. A., Morita, L., Kaufman, C., & Fedorenko, E. (2021). The neural basis of language processing: Transformer models as a unifying framework. *Proceedings of the National Academy of Sciences*, *118*(22), e2101372118. https://doi.org/10.1073/pnas.2101372118