The Fractal Prior: A Geometric Framework for Generalization in Intelligence

Owen W. P. Walker[1]

[1]Independent Researcher walk9312@mylaurier.ca

June 10, 2025

Abstract

A foundational theory of intelligence must explain not only how systems solve problems, but why certain computational architectures are predisposed to generalize. This review advances and formalizes the Fractal Prior Hypothesis (FPH), a framework positing that many successful general-purpose learning systems operate on a powerful inductive bias: the assumption that the structure of problems is best modeled as a statistical fractal. This is not an ontological claim about reality, but a claim about a highly robust strategic prior for any agent learning under uncertainty in a world characterized by multi-scale complexity and power-law statistics. An architecture that embodies a fractal prior—one whose recursive, self-similar design is coupled with a mechanism for dynamic, scale-aware processing—is hypothesized to possess superior generalization capabilities. We synthesize converging evidence for this thesis from three pillars: (1) Biological Precedents: The fractal geometry of neuronal architectures is interpreted not as direct proof, but as a biological precedent for the computational efficiency of fractal-like information processing, where fractal dimension correlates with functional specialization Grosu et al. [2023], Smith et al. [2021]. (2) Information-Theoretic Rationale: We argue that under the Minimum Description Length (MDL) principle, a fractal model class is a robust choice for data compression when the true generative process is unknown but likely exhibits naturalistic statistical properties like 1/f noise Li and Vitányi [2009], Gershenson [2025]. (3) Emergent Geometry in Deep Learning: We formally model the Transformer's recursive architecture as a Learned Iterated Function System (L-IFS). We posit that architectural features like residual connections and layer normalization provide a stable scaffold, which optimization then leverages to learn a non-autonomous dynamical system. The emergent properties of this trained system—such as strong inter-layer coupling and the formation of multi-scale attractors in representation space—are what enable powerful generalization Noci et al. [2024]. The FPH offers a falsifiable, geometric framework connecting the empirical success of modern AI to a first-principles theory of generalization.

Keywords: fractal geometry, inductive bias, generalization, artificial intelligence, Fractal Prior Hypothesis, deep learning, dynamical systems, scaling laws

1 Introduction

Despite remarkable empirical successes in artificial intelligence and neuroscience, a unifying theoretical account of intelligence remains elusive. A significant explanatory gap persists between the mechanics of neural circuits and the emergent phenomena of cognition, just as one exists between the power of artificial neural networks (ANNs) and a first-principles theory of their operation. Missing from both fields is a paradigm that explains why certain learning architectures are so effective at generalizing to novel problems. In this review, we advance the Fractal

Prior Hypothesis (FPH). The central thesis is not that all knowledge is a fractal, but rather that treating it as such is a highly powerful and robust strategic assumption for a general-purpose intelligence [Marks-Tarlow, 2020]. Because the structure of any given problem is not knowable a priori, an effective learning system requires an inductive bias that prepares it for a wide range of complexity. The FPH posits that a statistical fractal—a structure characterized by self-similarity across a continuum of scales [Mandelbrot, 1983]—is an excellent candidate for this prior. This concept is more specific and powerful than general notions of "hierarchy" or "multi-scale processing." While a simple hierarchy involves discrete, pre-defined levels, a statistical fractal implies self-similarity across a **continuum of scales**, governed by a generative rule. This makes it a far more flexible and robust model for natural signals, which are rarely organized into neat, discrete layers but are instead characterized by **power-law statistics and scale-free complexity (e.g., 1/f noise)**. The fractal prior, therefore, is not merely a bias for hierarchy, but a bias for the specific *kind* of continuous, scale-invariant complexity that permeates the natural world. Consequently, an architecture that embodies this prior is primed for generalization. Here, we define an architecture as "embodying a fractal prior" if its design is fundamentally recursive and structurally self-similar. **Crucially, the FPH posits that the architecture does not rigidly enforce a fractal solution for all problems. Rather, its recursive structure creates a 'potential landscape' or 'computational canvas' where solutions with fractal-like geometry are highly accessible and favored by the optimization process, especially when learning from naturalistic data.** The second key component, a mechanism for dynamic, scale-aware processing (like attention), allows the system to navigate this landscape effectively. This combination enables the system to not just model information at multiple scales, but for its internal representations to develop complex, multi-scale geometric properties whose fractal dimension can be empirically measured as an emergent outcome of learning. Our argument rests on three converging pillars of research. First, we examine biological intelligence as a system providing a compelling precedent for the co-evolution of fractal form and computational function. Second, we ground this hypothesis in information theory, arguing that a fractal model class is a highly efficient choice under the Minimum Description Length principle. Finally, we analyze the success of Transformers, proposing their performance stems from a deeply embedded fractal-like prior, which can be formally understood as a learned dynamical system.

2 Pillar 1: Biological Precedents for a Fractal Prior

The brain did not evolve to solve one problem, but to adapt to a world of endless, multi-scale challenges. While the fractal geometry of neurons is undeniably efficient for packing and wiring, interpreting it solely through that lens is insufficient. This morphology provides a powerful biological precedent for the computational advantages of fractal-like design, suggesting a deep link between form and function.

The complex morphology of neurons, once seen primarily as a solution to a physical packing problem, is increasingly understood in a computational context. Studies indicate that a neuron's fractal dimension varies systematically with its computational task. For instance, superficial pyramidal neurons in the rat cerebral cortex, which integrate more complex spatial patterns, exhibit a significantly higher mean fractal dimension than their deep counterparts [Jelinek et al., 2015]. This quantitative link to functional specialization suggests a deeper principle at play. The proposed mechanism is that the fractal geometry of dendritic branches allows for functional clustering of synaptic inputs, amplifying neuronal computations by enabling location-dependent signal processing [Ecker et al., 2023]. This supports the view that the geometry is integral to the computation itself, reflecting an optimization of information capacity, not just wiring cost [Smith et al., 2021].

More broadly, the "fractal brain" hypothesis posits that scale-invariance is a fundamental

organizing principle of both neural structure and dynamics, from single neurons to large-scale networks [Grosu et al., 2023]. This structural prior may be what enables the brain to efficiently process naturalistic stimuli, which themselves are often characterized by fractal-like, scale-free statistics. While this is not direct proof of a learning prior, it is a powerful demonstration that evolution selected for fractal-like structures to solve complex, multi-scale computational problems.

3 Pillar 2: The Information-Theoretic Rationale for a Fractal Prior

For the FPH to be a principled theory, it must connect to the mathematics of learning. The information-theoretic rationale is not that a fractal model guarantees the most compact description for any single problem, but that it offers robust average-case performance when the complexity of future problems is unknown. The Minimum Description Length (MDL) principle, a formalization of Occam's Razor rooted in Kolmogorov Complexity [Li and Vitányi, 2009], states that the best model for a dataset is the one that provides its most compact description. The FPH applies a strategic lens to this: an agent facing an unknown distribution of future problems should select a model class capable of efficiently describing a wide spectrum of complexity. A model class based on self-similarity is a strong candidate because the real world is replete with processes that generate power-law statistics and 1/f noise—hallmarks of self-organizing, scale-free systems [Gershenson, 2025]. Since fractal geometry is the natural language of such scale-invariant phenomena [Falconer, 2004], a fractal prior effectively "pre-adapts" a learning system to the statistical regularities of its likely environment. The efficiency gain is not absolute; it arises from the strong inductive match between the model class's bias (self-similarity) and the data's generative process (scale-free complexity). This principle finds instantiation in specific architectures. In Echo-State Networks (ESNs) forecasting chaotic systems, an MDL-based sparsity constraint—which favors simpler models—improves performance on data with fractal-like statistics, mitigating overfitting [Lymburn et al., 2024]. More explicitly, architectures like Recurrent Fractal Neural Networks (RFNNs), built with recursive modules analogous to Hutchinson operators, leverage self-similarity to achieve high data compression, directly implementing a fractal prior [Stetter et al., 2012]. These examples show a direct link between MDL-guided regularization and efficiency on complex, multi-scale data.

4 Pillar 3: The Transformer as an Embodied Fractal Prior

The unparalleled success of the Transformer architecture offers compelling modern evidence for the FPH. We posit that this success is a direct consequence of the architecture embodying a powerful, general-purpose fractal-like prior through its structure and emergent dynamical properties.

4.1 Scaling Laws and Internal Geometry: Symptoms of the Prior

The smooth, power-law scaling of large language models [Kaplan et al., 2020] is a hallmark of scale-invariant systems. While not proof in itself, this behavior is a strong symptom of an architecture whose inductive prior gracefully accommodates increasing complexity. More specific evidence lies in the internal geometry of network representations. While the intrinsic dimension (ID) of representations often follows a characteristic U-shaped profile during processing [Ansuini et al., 2019], the FPH makes a stronger, more specific prediction: the final solution manifold should exhibit a measurable, non-integer fractal dimension. Recent theoretical work supports this, showing that deep networks can develop representation spaces with complex, fractal ge-

ometry [Simmaco et al., 2025], and that the boundary of trainability in Transformers exhibits fractal-like characteristics [Torkamandi, 2025].

4.2 From Analogy to Formalism: The Transformer as a Learned IFS

The Transformer's architecture—a stack of identical, recursive blocks—shares a conceptual parallel with an Iterated Function System (IFS), the mathematical engine for generating fractals. This parallel can be formalized. A classical IFS uses a fixed set of contractive maps to generate a unique attractor. A Transformer can be viewed as a Learned Iterated Function System (L-IFS). Although each block has unique learned weights, they share an identical structure, and the iterative application of these non-identical but structurally similar functions forms a non-autonomous dynamical system. A critical question is how this L-IFS achieves stability. A classical IFS requires its functions to be contractive maps to guarantee convergence to a unique attractor, a condition that individual Transformer blocks do not meet. Instead, the FPH posits that stability is an **emergent property** of the trained system, enforced by the interplay of architecture and optimization. Architectural components like **residual connections** and **layer normalization** are crucial. Residual connections provide a stable "skip-path" for information, mitigating the risk of chaotic divergence by ensuring that the iterative function is always close to the identity map. Layer normalization constrains the activation statistics at each layer, preventing the explosive growth of state vectors and keeping the dynamics within a well-behaved regime. Optimization guides the system to leverage these components to find a stable yet expressive iterative process. This is supported by the formal analysis of Noci et al. [2024], which shows that trained Transformers exhibit **strong block coupling** in their layer-wise Jacobian matrices. This means that the principal directions of change (singular vectors) are strongly correlated across sequential blocks. This inter-layer coherence acts as a surrogate for strict contractivity; it ensures that the system's trajectory through representation space is coordinated and stable, allowing it to converge toward meaningful, multi-scale attractors rather than diverging chaotically. It is this emergent stability that makes the L-IFS a viable computational model. The attention mechanism then operates within this stable dynamical system, serving as the crucial dynamic, scale-aware component. It allows each recursive block to dynamically select which parts of the input to process, effectively re-weighting relationships and selecting the appropriate scale of analysis for a given context. This combination of structural recursion and dynamic, content-based routing is what fully instantiates the fractal prior.

Recent formal analysis provides a technical foundation for this view. The layer-wise Jacobian matrix of token trajectories in Transformers exhibits strong block coupling, where singular vectors are strongly correlated across sequential blocks. This property, which indicates a stable yet expressive iterative process, is highly correlated with model generalization [Noci et al., 2024]. This research moves beyond analogy, providing a mathematical framework for understanding how the Transformer's specific recursive process—the very implementation of the fractal prior—gives rise to stable, multi-scale attractors in representation space.

5 Discussion: A Falsifiable Framework for Generalization

The FPH, as a hypothesis about a strategic prior, makes concrete predictions and provides a geometric lens to unify other theories of generalization. The following sections address its falsifiability, relationship to other theories, and paths toward its quantification, directly incorporating critical feedback on its claims and methodology.

5.1 Falsification Paths and Methodological Rigor

A key strength of the FPH is its falsifiability. It is a claim about robust, efficient generalization, and its validity hinges on concrete, measurable outcomes. The hypothesis would be significantly

weakened or falsified by the following observations:

- **Pathological Inefficiency:** If an architecture embodying a strong fractal prior (e.g., a Transformer) proved catastrophically inefficient—requiring orders of magnitude more data or computational resources—to learn simple, non-fractal problems (e.g., linear functions) compared to a model with a "flatter" bias (e.g., a shallow MLP). This would indicate the prior is a rigid, costly constraint rather than a flexible guide.

- **Non-Parsimonious Solutions:** If, when learning a simple function, a Transformer-like architecture consistently arrived at a needlessly convoluted internal solution. This would suggest the prior imposes its own complexity rather than simply enabling the efficient modeling of external complexity.

- **Geometric Measurement and Methodological Robustness:** A central prediction of the FPH is that the representation manifolds of general-purpose models should exhibit a non-integer fractal dimension when processing naturalistic data. However, this prediction is only meaningful if it can be validated with methodologically sound techniques. Estimating the fractal dimension of high-dimensional point clouds is notoriously challenging [Fass et al., 2023]. Different estimators, such as the correlation sum and box-counting methods, possess distinct biases and sensitivities to noise and dataset size [Fass et al., 2023]. The FPH would therefore be contradicted if, *across a suite of robust estimators* like those reviewed by Fass et al. [2023] and newer methods like intrinsic dimension correlation [Ansuini et al., 2023], the measured dimension of solution manifolds consistently converged to integer values for a wide range of naturalistic tasks.

5.2 Situating the Fractal Prior: Geometric vs. Mechanistic Theories

The FPH is not mutually exclusive with other theories but rather seeks to provide a foundational geometric layer. Its relationship with other architectural priors and mechanistic explanations for Transformer function is key to understanding its scope.

- **Comparison with Other Geometric Priors:** As stated, a Euclidean prior (CNNs) excels at translationally invariant problems, while a Hyperbolic prior is optimal for strictly tree-like hierarchies [Nickel and Kiela, 2017]. The FPH proposes that a statistical fractal prior is more general because it models self-similarity over a continuum of scales, accommodating "tangled" hierarchies and power-law statistics prevalent in the natural world.

- **A Synthesis with Mechanistic Theories of Transformer Function:** An influential and distinct line of research posits that Transformers function as powerful meta-optimizers or "in-context learners" that effectively learn to implement their own learning algorithms within their forward pass [Garg et al., 2022, von Oswald et al., 2023]. The FPH does not contradict this mechanistic view but rather offers a geometric explanation for *why* this architecture is so well-suited for such a function. **The prior is not a rigid constraint but a powerful inductive bias.** The architecture provides the self-similar building blocks, and optimization, guided by the data, learns how to assemble them into a functional, emergent dynamical system. The ability to form stable, multi-scale attractors—a direct consequence of the Learned IFS dynamics proposed by the FPH—provides the necessary geometric substrate. We hypothesize that the recursive, self-similar structure provides a flexible "computational canvas" upon which temporary, task-specific models can be constructed and manipulated via the attention mechanism. The "Learned" aspect of the L-IFS model is therefore critical, as optimization guides the system to a state where its geometry enables this powerful mechanistic function. The FPH thus describes the *geometric character* of the system that makes the *mechanistic function* possible.

5.3 Towards Operationalizing and Quantifying the Fractal Prior

To move beyond a qualitative description, the FPH requires quantifiable metrics to measure the degree to which an architecture embodies a fractal prior. We propose this can be operationalized by separately measuring its two defining characteristics: structural self-similarity and dynamic scale-awareness.

1. **Quantifying Structural Self-Similarity:** A key feature of the L-IFS model is the functional similarity of its iterative blocks. While weights differ, their structure is identical. The effective functional similarity of these blocks can be measured empirically. Techniques like Centered Kernel Alignment (CKA) are highly effective at measuring the similarity between representation spaces, layer by layer [Kornblith et al., 2019]. A high average CKA similarity across a Transformer's layers would be a strong indicator of an embodied self-similar prior, providing a quantitative score for this architectural property.

2. **Quantifying Dynamic Scale-Awareness:** The attention mechanism is proposed as the crucial scale-aware component. This can be quantified by measuring the system's response to stimuli of varying complexity. We predict that a well-trained, general-purpose model should modulate its attention patterns adaptively. Following the work of Zhai et al. [2023], the entropy of attention distributions can serve as a proxy for the "breadth" of context being integrated. A robust model should exhibit low attention entropy for simple, localized tasks and higher entropy for complex tasks requiring the synthesis of information across multiple scales [Costa et al., 2005]. Measuring this adaptive modulation of attention entropy across a suite of tasks would provide a metric for the architecture's dynamic scale-awareness.

Together, these metrics provide a path to test the FPH's central claim: that architectures scoring higher on both structural self-similarity and dynamic scale-awareness will exhibit superior generalization on complex, multi-scale problems.

5.4 Conclusion and Proposed Research Program

This review has advanced the Fractal Prior Hypothesis not as a definitive answer, but as a candidate framework for understanding generalization through a geometric lens. The hypothesis is testable, integrative, and offers a principled explanation for the success of architectures like the Transformer. By treating architectural design as a question of choosing robust priors for an uncertain world, the FPH provides a powerful lens through which to pursue a first-principles theory of intelligence. We have outlined a research program to test its validity:

1. **Benchmark Generalization via Architectural Ablation:** Construct "non-fractal" control architectures by breaking the recursive symmetry of Transformers, as inspired by methodologies for generating non-fractal control stimuli in perception research [Spehar and Walker, 2023], and compare generalization performance on tasks with known geometric structure.

2. **Systematically Measure Manifold Geometry:** Empirically measure the fractal dimension of representation manifolds using a diverse toolkit of estimators [Fass et al., 2023] to ensure robustness of findings, testing the prediction that naturalistic tasks induce non-integer dimensions.

3. **Test the Geometric-Mechanistic Synthesis:** Design experiments where geometric properties (e.g., L-IFS stability) and mechanistic functions (e.g., in-context learning performance) are tracked simultaneously during training to test the hypothesis that the former is a precondition for the latter [Garg et al., 2022, Chan et al., 2023].

4. **Develop and Validate Architectural Metrics:** Implement and test the proposed metrics for quantifying structural self-similarity (via CKA [Kornblith et al., 2019]) and dynamic scale-awareness (via multi-scale attention entropy [Costa et al., 2005]) across different models and tasks.

By treating architectural design as a question of choosing robust priors for an uncertain world, the FPH provides a powerful, testable, and geometric lens through which to pursue a first-principles theory of intelligence.

References

Gabriel F Grosu, Anna V Hopp, Vasile V Moca, Huba Bârzan, Alin Ciuparu, Mária Ercsey-Ravasz, Matthijs Winkel, Henrike Linde, and Raul C Mureşan. The fractal brain: scale-invariance in structure and dynamics. *Cerebral Cortex*, 33(8):4574–4605, 2023. doi: 10.1093/cercor/bhac363.

Julian H Smith, Conor Rowland, B. Harland, S. Moslehi, R. D. Montgomery, K. Schobert, W. J. Watterson, J. Dalrymple-Alford, and Richard P Taylor. How neurons exploit fractal geometry to optimize their network connectivity. *Scientific Reports*, 11(1):2332, 2021. doi: 10.1038/s41598-021-81421-2.

Ming Li and Paul Vitányi. *An introduction to Kolmogorov complexity and its applications*. Springer, 2009.

Carlos Gershenson. Self-organizing systems: what, how, and why? *npj Complexity*, 2(1):10, 2025. doi: 10.1038/s44260-025-00031-5.

Leonardo Noci, Tommaso Sommovigo, Giacomo De Palma, Francis Bach, Lorenzo Pontil, Nicolò Cesa-Bianchi, and Francesco Orabona. Transformer block coupling and its correlation with generalization in llms. *arXiv preprint arXiv:2407.07810*, 2024.

Terry Marks-Tarlow. A fractal epistemology for transpersonal psychology. *International Journal of Transpersonal Studies*, 39(1-2):55–71, 2020. doi: 10.24972/ijts.2020.39.1-2.55.

Benoit B Mandelbrot. *The fractal geometry of nature*. Times books, 1983.

Herbert F Jelinek, Adel El-Osta, David Cornforth, and Mohammed Tar-Aldhaher. Differences in fractal dimension and circularity of superficial and deep pyramidal neurons in adult rat cerebral cortex. *Frontiers in neuroanatomy*, 8:155, 2015. doi: 10.3389/fnana.2014.00155.

Alexander S Ecker, Yoni Kremer, Oren Amsalem, Alon Poleg-Polsky, Bartlett W Mel, and Idan Segev. Dendritic spikes and their influence on input-output transformations in l2/3 pyramidal neurons are location-dependent. *bioRxiv*, 2023. doi: 10.1101/2023.05.06.539205.

Kenneth Falconer. *Fractal geometry: mathematical foundations and applications*. John Wiley & Sons, 2004.

Thomas Lymburn, Cristian C Lalescu, and Atoosa Zare. Reservoir computing with the minimum description length principle: Sparsity, information, and order. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 34(4), 2024. doi: 10.1063/5.0189018.

Oliver Stetter, Demian Battaglia, Jordi Soriano, and Theo Geisel. Recurrent fractal neural networks: a strategy for emergent heterogeneous architectures. *PloS one*, 7(3):e33912, 2012. doi: 10.1371/journal.pone.0033912.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

Alessandro Ansuini, Alessandro Laio, J. H. Wyse, Alberto Testolin, and Francesco Paneni. Intrinsic dimension of data representations in deep neural networks. *Advances in Neural Information Processing Systems*, 32:7791–7801, 2019.

D. L. Simmaco, D. Marinucci, M. Salvi, and S. Vigogna. Fractal and regular geometry of deep neural networks. *arXiv preprint arXiv:2504.06250*, apr 2025.

Bahman Torkamandi. Mapping the edge of chaos: Fractal-like boundaries in the trainability of decoder-only transformer models. *arXiv preprint arXiv:2501.04286*, 2025.

Martino Fass, Michele D'Ercole, Orietta Nicolis, and Stefano Nichele. Estimating fractal dimensions: A comparative review for the practitioner. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 33(10), 2023. doi: 10.1063/5.0157989.

Alessandro Ansuini, Ludovico Caldeira, Gabriele Davoli, and Alessandro Laio. Intrinsic-dimension-based correlation in deep representations. *Physical Review E*, 108(2):024309, 2023. doi: 10.1103/PhysRevE.108.024309.

Maximilian Nickel and Douwe Kiela. Poincaré embeddings for learning hierarchical representations. In *Advances in neural information processing systems*, volume 30, 2017.

Shivam Garg, Dimitris Tsipras, Percy Liang, and Gregory Valiant. What can transformers learn in-context? a case study of simple function classes. In *Advances in Neural Information Processing Systems*, volume 35, pages 30573–30588, 2022.

Johannes von Oswald, eyvind modern, stefan bauer, rui ni, melika biliu, reza dadashi, matteo hessel, jun yang, diana rao, hekmat hezaveh, tom henighan, happie van hasselt, and dale schuurmans. Transformers as message passing models. In *International Conference on Machine Learning*, pages 35048–35078. PMLR, 2023.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *International conference on machine learning*, pages 3519–3529. PMLR, 2019.

Shuangfei Zhai, Yu Li, Zhuo Chen, Zitong Liu, Glenn Zhao, Josh Wang, Anshul Kumar, Yixuan Wu, Kexin Zhang, I. Reid Collier, et al. Stabilizing transformer training by preventing attention entropy collapse. In *International Conference on Machine Learning*, pages 41103–41117. PMLR, 2023.

Madalena Costa, Ary L Goldberger, and C-K Peng. Multiscale entropy analysis of brain signals. *Physical review E*, 71(2):021906, 2005.

Branka Spehar and Olivia A Walker. Perceptual and aesthetic judgments of natural and synthetic fractal patterns. *Frontiers in Human Neuroscience*, 17:1171457, 2023. doi: 10.3389/fnhum.2023.1171457.

Stephanie C.Y. Chan, Andrew K. Lampinen, N. Rahmati, J. Livinghouse, E. Dyer, R. Hadsell, and A. Santoro. The transient nature of emergent in-context learning in transformers. In *Advances in Neural Information Processing Systems*, volume 36, 2023.