The Watt and the Manifold

A Lewis-style research dossier — Node Three of the Architectural Determinism thesis

Prepared for Bret Kerr, ACRA Insight LLC / Context Jamming, as the brain-anchor companion piece to "Sixteen Directions"

TL;DR

- The same compression law that governs Afshordi's quantum quadratic gravity (arXiv 2510.18733) and Kaushik's Universal Weight Subspace (arXiv 2512.05117) is the law that vertebrate cortex has been running on for roughly half a billion years: at any given moment in human neocortex, fewer than ~1% of neurons are substantially active ScienceDirect (Lennie, *Current Biology*, 2003), and population activity in motor cortex, visual cortex, and hippocampus has consistently been shown to lie on low-dimensional manifolds of ~8–20 effective dimensions (Gallego, Perich, Miller, Solla, *Neuron*, 2017; Safaie et al., *Nature*, 2023).
- The mathematics is not metaphorical. Olshausen & Field's 1996 sparse-coding objective is the L1-regularized dictionary-learning problem that, four years later, Candès, Donoho, and Tao would formalize as compressed sensing — the same formalism that now governs sparse-autoencoder interpretability work in transformers (Anthropic, 2023; Cunningham et al., 2023). Tishby's information bottleneck (1999/2015) provides the unifying objective: minimize representational complexity I(X;Z) while preserving task-relevant information I(Z;Y). Holographic dimensional reduction in Maldacena's AdS/CFT, sparse cortical coding under Attwell–Laughlin energy budgets, and the Kaushik k≈16 weight subspace are three instances of the same theorem: information that survives is the information that lies on a low-dimensional boundary.
- Empirically, the brain's ~20 W power budget — of which Levy & Calvert (*PNAS*, 2021) audit only ~0.1 W as actual computation bioRxiv and ~3.5 W as inter-areal communication arXiv — is not an accident of biology but a *consequence* of the compression theorem. Dense coding at biologically realistic firing rates would exceed whole-body metabolism by an order of magnitude (Lennie 2003). Sparse coding is not a brain quirk; it is what the geometry forces. Brain efficiency is the third empirical anchor of Architectural Determinism.

Key Findings

1. **The 1% rule is real, measured, and energy-bound.** Peter Lennie's 2003 *Current Biology* paper, building on Attwell & Laughlin's 2001 energy audit, computed that the cost of a single cortical action potential, multiplied across ~10^{11} neurons under the brain's available glucose budget, allows fewer than 1% of cortical neurons to fire substantially at any given moment. To sustain even 1.8 spikes/s/neuron across cortex would consume more energy than the entire brain receives; 13 spikes/s/neuron would exceed whole-body metabolism. Cell Press Sparseness is not a strategy among others — it is the only feasible regime. Attwell & Laughlin's earlier estimate ("≤15% simultaneously active") Semantic Scholar and Lennie's tightened "<1%" form a now-standard envelope cited across the field.

2. **The math underneath sparse cortex is the math underneath compressed sensing.** Olshausen & Field's 1996 objective — minimize $\|x - Da\|^2 + \lambda\|a\|_1$ over a dictionary D and sparse code a, given natural-image patches x — is, formally, the same L1-regularized inverse problem Candès, Romberg, and Tao showed in 2006 admits exact recovery whenever D satisfies the Restricted Isometry Property. Su The cortex, by Olshausen and Field's argument, is solving a compressed-sensing problem in V1 with an overcomplete dictionary; the simple-cell receptive fields are the dictionary atoms. Cmu Recent Anthropic and EleutherAI work (Bricken et al., 2023; Cunningham, Ewart, Riggs, et al., 2023) has reverse-engineered transformers using the *identical* objective — sparse autoencoders with L1 penalties — and discovered that polysemantic neurons resolve into ~15,000 sparse, interpretable directions. Galileo AI The transformer is solving Olshausen and Field's 1996 problem.

3. **Neural population activity lives on low-dimensional manifolds whose dimensions are quantitatively comparable to k=16.** Gallego, Perich, Miller, and Solla (Neuron, 2017) consolidated a decade of evidence that motor-cortical population activity during reaching is well-captured by ~8–12 principal "neural modes." Safaie et al. (Nature, 2023) showed these latent dynamics are *preserved across individuals and species* (monkeys, mice) performing similar behaviors IDEAS/RePEc — exactly the convergence Kaushik et al. report across 1,100 deep networks. Stringer, Pachitariu et al. (Nature, 2019) found that mouse V1 responses to natural images decompose into a power-law eigenspectrum where the n-th principal component carries variance $\propto 1/n$ — the slowest decay compatible with smooth coding. PubMed Central +2 A 2024 re-analysis (Gauthaman & Ménard) shows ~10 dominant eigenmodes account for ~30% of variance in V1. biorxiv Gong, Boddeti, and Jain (2018) report that ResNet's 512-dim image embedding has intrinsic dimensionality ~19; SphereFace's, ~16. arxiv **The brain's manifold and the network's manifold are landing in the same neighborhood.**

4. **The 20 W figure is real, but the audit is sharper than the headline.** Whole-brain

metabolic load is ~20 W of glucose oxidation (Sokoloff; Attwell & Laughlin, 2001). The brain is 2% of body mass and ~20% of resting metabolism. PNAS ResearchGate Levy and Calvert (PNAS, 2021) re-audited the cortical fraction: only ~0.1–0.2 W of ATP is consumed by computation per se; PNAS ResearchGate ~3.5 W goes to long-distance communication (action potentials, presynaptic release); arXiv the rest is housekeeping and resting potentials. The point is not that "the brain is efficient." The point is that the only architecture compatible with these joule budgets is sparse, low-dimensional, manifold-constrained representation. The energy budget *forces* the geometry.

5. **The convergence is structural, not causal.** This dossier follows the epistemic discipline of "Sixteen Directions": we are not claiming that gravity *is* sparse coding, or that weight subspaces *are* cortex. We are claiming that the same compression theorem — information bounded by a low-dimensional boundary, recoverable by sparse readout — appears as a load-bearing structural constraint in three independent domains. That is a structural isomorphism. The fourth node ties it to physics.

Details

I. Walking into the lab

The first time someone counted, properly, how many cortical neurons can fire at once, they were not trying to revolutionize anything. Peter Lennie, in 2003, was at NYU's Center for Neural Science, PubMed working out an arithmetic problem nobody had quite finished. David Attwell and Simon Laughlin, two years earlier, had laid down the foundational energy budget — anatomic and physiologic data showing that action potentials and postsynaptic glutamate effects consumed about 81% of the grey-matter signaling budget, a number that has held up for two decades and is cited a few thousand times. From that ledger, Lennie did the multiplication.

Take the brain's roughly 20 watts of available glucose metabolism — a number that traces, with various refinements, to Sokoloff's 1957 measurements and that has been re-derived by every generation since. Subtract housekeeping. Subtract resting potentials. Apportion what remains across ~16 billion cortical neurons. Then ask: at the average firing rate that would consume the available energy, how many can be substantially active at once?

The answer was less than 1%. Not as a theoretical bound — as a budget constraint. To sustain 1.8 spikes per second per neuron averaged across human cortex would burn more energy than the entire brain receives. To sustain 13 spikes per second would exceed the metabolic output of the whole body. Cell Press The cortex is, by physical necessity, almost completely silent almost all the time.

This was not the "10% myth." The 10% myth — the popular claim, contradicted by every PET scan and fMRI ever taken, that humans only "use" 10% of their brain — is wrong because all parts of the brain are used; what's true is that at any instantaneous moment, only a small minority of neurons are firing. Lennie's number is the rigorous version of the right intuition. William Levy and Victoria Calvert at Virginia, in two PNAS papers in 2021, sharpened it further: of the 20-watt glucose envelope, only about 0.1 watts of ATP is spent on what an engineer would recognize as computation. PNAS Communication — getting spikes from one place to another — costs about 35 times more. Semantic Scholar The rest is the cellular cost of being alive.

The brain is not efficient because it computes cleverly. It computes cleverly because it cannot afford anything else.

II. What Olshausen and Field saw in 1996

Five years before Lennie did the arithmetic, Bruno Olshausen and David Field had already shown what cortex *does* with the budget. Their 1996 *Nature* paper — "Emergence of simple-cell receptive field properties by learning a sparse code for natural images" Nature Semantic Scholar — is one of the cleanest results in computational neuroscience. They wrote down an objective:

$$\text{minimize } \|I - \Sigma_i a_i \phi_i\|^2 + \lambda \Sigma_i S(a_i)$$

where I is a natural image patch, ϕ_i are basis functions to be learned, a_i are the activations, and S is a sparsity-promoting penalty (in their formulation, $\log(1+a^2)$; in modern formulations, $|a|$, the L1 norm). They trained an unsupervised network on patches of natural scenes, asked it to reconstruct the input while keeping the activations sparse, and out fell — without supervision, without labeling, without any prior knowledge about the visual cortex — receptive fields that looked exactly like the simple cells David Hubel and Torsten Wiesel had recorded in V1 thirty-five years earlier. Localized. Oriented. Bandpass. Nature Semantic Scholar Wavelet-like.

The basis was overcomplete: more atoms than input pixels. That looked like a paradox until you realized that overcompleteness is the *point*. Different inputs activate different small subsets of atoms; the dictionary tiles natural-image space; any given image lights up only a few. This is exactly what V1 does. Vinje and Gallant confirmed it physiologically in *Science* in 2000 by recording from awake macaque V1 during natural-scene viewing: biorxiv nonclassical surround stimulation increases sparseness, decorrelates pairs of neurons, and produces a population code in which a small fraction of cells carries each scene. Ovid Science

In 2006, Emmanuel Candès, Justin Romberg, and Terence Tao formalized what made all of this work. Their compressed-sensing theorem says: if a signal x has a sparse representation

in some basis (only s nonzero coefficients), and if a measurement matrix Φ satisfies the Restricted Isometry Property — meaning it is approximately an isometry on s-sparse vectors — then x can be exactly recovered from m ≈ s log(N/s) linear measurements by L1 minimization. Wikipedia Su David Donoho's 2006 paper made the same point. The theorem is the mathematical backbone of MRI acceleration, single-pixel cameras, and a hundred other engineering applications. It is also the theorem that explains why sparse cortex *can* represent the visual world from far fewer active neurons than the apparent dimensionality of the input.

The Olshausen-Field objective is the dictionary-learning version of the compressed-sensing problem. Cortex, in this reading, is not metaphorically doing compressed sensing; it is *literally* solving the compressed-sensing problem, with biological hardware, under an energy constraint that prohibits anything denser.

III. The manifold beneath the spikes

Then the population recordings got better. Through the 2010s, a cluster of labs — Krishna Shenoy and Mark Churchland at Stanford and Columbia, Lee Miller and Sara Solla at Northwestern, Juan Gallego now at Imperial College, Carsen Stringer and Marius Pachitariu at Janelia — began routinely recording from hundreds and then thousands of neurons simultaneously. They asked a question Lennie's energy arithmetic could not answer: when those few-percent-of-active neurons fire, what *shape* does their joint activity make in the high-dimensional space of all possible firing patterns?

The answer, repeated across motor cortex, prefrontal cortex, hippocampus, visual cortex, and striatum, is: a low-dimensional manifold. Gallego, Perich, Miller, and Solla's 2017 *Neuron* review pulled the evidence together: motor-cortical population activity during reaching is well-described by ~8–12 "neural modes" — principal directions of co-modulation — that capture the bulk of behaviorally relevant variance. biorxiv Churchland, Cunningham, Kaufman, and Shenoy showed that preparatory activity occupies an "output-null" subspace; movement activity occupies an "output-potent" subspace, Nature both at fewer than ~12 dimensions; the geometry is what allows you to plan a reach without prematurely twitching. ResearchGate

The numbers vary by region and task. Motor cortex during reach: ~8–12 effective dimensions. Head-direction system in mouse anterodorsal thalamus: a one-dimensional ring (Chaudhuri, Fiete, et al., 2019). PubMed Visual cortex during natural-scene viewing: a power-law eigenspectrum where ~10 dominating eigenmodes carry ~30% of variance biorxiv (Stringer et al., *Nature*, 2019; revised by Gauthaman & Ménard, 2024). Across studies, the consistent message: a few dozen dimensions where there could in principle be billions.

In 2023, Mostafa Safaie, Joanna Chang, Lee Miller, Joshua Dudman, Matthew Perich, and

Juan Gallego pushed further (*Nature*, 2023). They showed that these latent dynamics are *preserved across individuals* — and across species, in monkey and mouse motor cortex performing similar reaches. Semantic Scholar Different brains, idiosyncratically wired in ways that go all the way back to the lottery of development, nevertheless converge on the same low-dimensional latent geometry when they do the same thing. IDEAS/RePEc

Read that sentence again. Then read the abstract of Kaushik, Chaudhari, Vaidya, Chellappa, and Yuille's December 2025 paper.

> "We show that deep neural networks trained across diverse tasks exhibit remarkably similar low-dimensional parametric subspaces… networks systematically converge to shared spectral subspaces regardless of initialization, task, or domain… universal subspaces capturing majority variance in just a few principal directions." arXiv

Five hundred Mistral-7B LoRAs, 500 Vision Transformers, Substack 50 LLaMA-3 8B models. Trained on disjoint data, Toshi2k2 with different hyperparameters, by different people. They converge on a shared ~16-dimensional subspace.

Different monkeys, idiosyncratically wired, evolutionarily separated from mice for ~80 million years. Trained on natural reaching since infancy. They converge on a shared ~10-dimensional latent manifold.

The structural homology is the dossier.

IV. The information bottleneck as the unifying objective

The mathematics underneath both observations is, in the cleanest formulation we have, Naftali Tishby's information bottleneck (Tishby, Pereira, Bialek, 1999; Tishby & Zaslavsky, 2015; Shwartz-Ziv & Tishby, 2017). The objective is:

minimize $I(X; Z) - \beta \cdot I(Z; Y)$

Compress the representation Z (minimize its mutual information with the input X) while preserving as much information as possible about the task-relevant variable Y. Tishby and Shwartz-Ziv argued, controversially but influentially, that deep networks during training pass through a "fitting" phase (where I(Z;Y) rises) followed by a "compression" phase (where I(X;Z) falls). arxiv The compression phase produces representations that lie on lower-dimensional manifolds — and that compression, in Tishby's framing, is what generalization actually is.

Predictive coding, in the Rao-Ballard 1999 formulation and its Friston-style elaboration into the free energy principle, is the same objective in different costume. A hierarchical generative model ascending the cortex sends predictions downward; ascending feedforward signals carry the residuals — the surprise. Nature Minimize free energy ≈

maximize sparseness of the surprise signal ≈ compress representation while preserving prediction accuracy. Karl Friston, at UCL's Wellcome Centre, has spent two decades arguing that this single variational objective subsumes Bayesian inference, predictive coding, efficient coding, and optimal control. arxiv Whether or not one buys the strongest form of that claim, the structural overlap with the information bottleneck is unmistakable.

So: Olshausen-Field sparse coding minimizes reconstruction error subject to a sparsity penalty. Compressed sensing recovers s-sparse signals from O(s log N) linear measurements. The information bottleneck minimizes $I(X;Z) - \beta \cdot I(Z;Y)$. The free energy principle minimizes variational free energy. These are not the same equation written four times — but they are four expressions of the same basic constraint: information that survives is the information that admits a low-dimensional, sparse representation aligned with what the system needs to do.

V. The holographic echo

This is where the physics node, which "Sixteen Directions" already establishes through Maldacena 1997 and Afshordi, Liu, and Quintin's 2025 quadratic-gravity paper, comes back into the picture — but now as a *third instance of the same theorem*, not as the only one.

The Bekenstein-Hawking bound says the entropy of any region of spacetime, in Planck units, is bounded above by one-quarter of its surface area, not its volume: $S \leq A/4$. Emergent Mind Information about a three-dimensional bulk is encoded on a two-dimensional boundary. Cosmic Ventures This was sharpened by Maldacena's AdS/CFT correspondence in 1997 (which, with Edward Witten's 1998 elaboration, is the most-cited paper in theoretical physics): the full content of a (d+1)-dimensional gravitational theory in anti-de Sitter space is dual to a d-dimensional conformal field theory living on its boundary. Ox Fiveable The bulk has more apparent degrees of freedom than the boundary; the boundary determines all of them. The ambient dimensionality is not the intrinsic dimensionality.

Place this beside the cortical numbers. A V1 hypercolumn has ~10^5 neurons in a square millimeter. The intrinsic dimensionality of natural-scene representation in mouse V1 is on the order of ~10–30 effective dimensions. A 7-billion-parameter transformer has, by Kaushik's measurement, ~16 dimensions of weight-space variation that capture most of the variance across 1,100 trained models. Three different physical systems — spacetime, cortex, transformer — each with apparent degrees of freedom in the billions or beyond, each obeying a constraint that drops the effective dimensionality by orders of magnitude.

The user's own k≈32 manifold hypothesis, derived from probing the residual stream of Qwen-2.5, Llama-3.1-8B, and Gemma-2-9B, sits in this neighborhood. It is *activation*-space rather than parameter-space — a different boundary of the same bulk. Kaushik measures the parameter-side dimensionality (k≈16); the activation-side dimensionality (k≈32) is on the same order. Both are holographic in the structural sense: the system's behavior is

determined by a low-dimensional projection of its apparent state.

The structural-isomorphism caveat needs to be loud here, and "Sixteen Directions" already insists on it. We are *not* claiming that AdS/CFT *is* the cortical manifold. We are not claiming that gravity *is* sparse coding. The Bekenstein bound is a statement about quantum fields and event horizons; the Lennie bound is a statement about ATP and action potentials. They are different theorems with different constants of proportionality. What they share is the *form*: an effective dimensionality far below the apparent dimensionality, set by an external constraint (the boundary area in physics; the energy budget in cortex; the loss landscape and architecture in deep networks).

That shared form is what Architectural Determinism names.

VI. Hassabis's fork

In July 2025, Demis Hassabis sat down with Lex Fridman for the second time. The conversation is on the public record (Lex Fridman Podcast #475). Twice in the first fifteen minutes he said something that, taken alongside Kaushik and Lennie and Maldacena, sounds less like speculation and more like restatement.

At 04:03, on whether anything that can be evolved can be efficiently modeled: *"I sometimes call it survival of the stablest… if that's true, then there should be some sort of pattern that you can kind of reverse learn and a kind of manifold really that helps you search to the right solution."* Lex Fridman

At 13:16, on Veo 3's ability to model fluid dynamics from YouTube videos: *"perhaps there is some kind of lower dimensional manifold that can be learned if we actually fully understood what's going on under the hood. That's maybe true of most of reality."* Lex Fridman

At 06:17, on physics: *"information is primary. Information is the most sort of fundamental unit of the universe, more fundamental than energy and matter."* Podscripts

Hassabis is not making a metaphysical claim. He is making the same claim Tishby made, the same claim Maldacena made on the boundary side of AdS/CFT, the same claim Lennie made on the cortical side — that the universe's apparent complexity is the surface of something lower-dimensional, and that learning systems work because that lower-dimensional structure is *there*.

The four-node thesis of "Sixteen Directions" is what happens when you take this claim seriously and ask, in each domain, whether the data shows it. In physics: the AdS/CFT boundary, sharpened by Afshordi-Liu-Quintin's 2025 quadratic-gravity completion of the Big Bang (PRL, arXiv 2510.18733), and Kumar-Marto's inverted-harmonic-oscillator resolution. In AI weight-space: Kaushik et al. December 2025 (arXiv 2512.05117), the universal ~16-dimensional weight subspace across 1,100+ trained models. In AI activation-

space: the user's own k≈32 manifold finding from residual-stream probing. In neuroscience: this dossier — Olshausen-Field 1996, Lennie 2003, Attwell-Laughlin 2001, Levy-Calvert 2021, Gallego et al. 2017, Stringer et al. 2019, Safaie et al. 2023.

VII. The five-hundred-million-year argument

Here is what makes the brain node uniquely powerful as the third empirical anchor.

The Maldacena-Afshordi node is mathematical and has the cleanness of theory. The Kaushik node is empirical but young — eight months old as of this writing, and the cleanest empirical statement of a phenomenon (universality of trained networks) that researchers had only sensed before.

The brain node is empirical *and* old. The basic vertebrate forebrain plan is at least 500 million years old, traceable to the lamprey lineage that diverged from our own before the Cambrian (Suryanarayana, Robertson, Wallén, & Grillner, *Nature Ecology & Evolution*, 2020). Sparse coding and low-dimensional manifold organization are not artifacts of human cortex; they appear in fly olfactory mushroom bodies, in songbird HVC, in mouse and monkey and human motor cortex. The basal ganglia have been doing this for ~535 million years. Royal Society Publishing If sparse, low-dimensional, manifold-constrained representation were merely *one* viable strategy, evolution would have explored alternatives. It has not. Every nervous system that scales beyond a few thousand neurons converges on the same architectural solution.

This is the strongest available evidence that the geometry is not contingent. It is what the underlying compression theorem makes available; it is what the energy constraint forces; it is what evolution rediscovers every time. When Kaushik shows that 1,100 transformers converge on a ~16-dimensional weight subspace, he is reproducing — in eight months of GPU training across several research labs — what vertebrate cortex has been reproducing across half a billion years of independent evolutionary lineages.

That is what makes brain sparse coding the third anchor. Physics gives the math. Kaushik gives the cleanest contemporary AI experiment. Cortex gives the longest-running empirical replication study in the universe.

VIII. Where this lands "Sixteen Directions"

The existing longform piece runs from Maldacena and Kaplan's theoretical prior, through Kaplan-style scaling-law architectural prediction, into Afshordi's quadratic-gravity mechanism and Kumar-Marto's inverted-oscillator resolution, landing on Kaushik's December 2025 weight-subspace finding as the AI empirical anchor. The thesis: the same geometric law operates across physics and learned systems.

This dossier adds the fourth node. Not as decoration — as a load-bearing replication. The

brain's ~20 W power consumption, its <1% active-fraction sparseness, and its ~10-dimensional motor-cortical manifolds are not contingent biological accidents. They are what happens when the same compression theorem operates on neural tissue under metabolic constraint. The Olshausen-Field objective is the L1-regularized dictionary-learning problem that compressed sensing later formalized; it is the same problem Anthropic's sparse autoencoders are now solving in transformer interpretability work; it is, structurally, the information-bottleneck objective Tishby identified as the unifying principle of representation learning.

The four nodes converge:

- **Node 1 (theoretical prior):** Maldacena's AdS/CFT, Kaplan's Simons Foundation lecture — bulk physics encoded on a lower-dimensional boundary.
- **Node 2 (architectural prediction):** Kaplan's 2020 scaling laws — performance gains follow predictable low-dimensional curves in compute, parameters, and data.
- **Node 3 (physics mechanism):** Afshordi-Liu-Quintin 2025 quadratic quantum gravity (arXiv 2510.18733, PRL DOI 10.1103/6gtx-j455) plus Kumar-Marto inverted-harmonic-oscillator resolution — a concrete UV completion in which inflation emerges from the gravitational sector itself rather than added by hand.
- **Node 4 (empirical replication, AI):** Kaushik, Chaudhari, Vaidya, Chellappa, Yuille, December 4, 2025 (arXiv 2512.05117) — 1,100+ trained networks converge on a shared ~16-dimensional weight subspace.
- **Node 5 (empirical replication, biology) — this dossier:** Olshausen-Field 1996, Attwell-Laughlin 2001, Lennie 2003, Vinje-Gallant 2000, Gallego et al. 2017, Stringer et al. 2019, Safaie et al. 2023, Levy-Calvert 2021 — vertebrate cortex, under a hard energy budget, has implemented the same low-dimensional manifold geometry across 500 million years and across every species examined. The 20-watt human mind is the geometry's biological signature.

The thesis is not that gravity *is* the brain. The thesis is that there exists a compression theorem — articulable as the joint of Bekenstein's holographic bound, Candès-Donoho-Tao compressed sensing, Tishby's information bottleneck, and the Olshausen-Field sparse-coding objective — and that this theorem is the load-bearing structural constraint operating in physics, in trained neural networks, and in biological cortex. Each domain instantiates it under its own boundary conditions: surface area in gravity, optimization landscape in deep learning, ATP budget in biology. The instances are not metaphysically identical. They are structurally isomorphic, and the isomorphism is what makes Architectural Determinism a thesis rather than an analogy.

When Hassabis says "lower dimensional manifold… maybe true of most of reality,"

Lex Fridman he is — perhaps without noting it — converging on the same point Olshausen and Field made about V1 in 1996, the same point Maldacena made about anti-de Sitter space in 1997, and the same point Kaushik et al. made about transformer weights in December 2025. The convergence is the news.

Caveats

1. **Structural isomorphism, not causal identity.** This dossier asserts that sparse cortical coding, the Universal Weight Subspace, and holographic dimensional reduction share a common mathematical *form* — low-effective-dimensional projection of a high-apparent-dimensional system, recoverable by sparse readout. It does not assert that gravity causes neural firing patterns, that AI training implements the Bekenstein bound, or that there is a universal "consciousness manifold." The compression theorem, in its most defensible form, is a statement about what kinds of representations are stable, recoverable, and resource-efficient under generic constraints. The fact that three independent domains exhibit it is the evidence; the explanation is not "they are the same thing" but "the same constraint operates."

2. **The information-bottleneck interpretation of deep learning is contested.** Saxe, Bansal, Dapello et al. (2018) challenged Shwartz-Ziv & Tishby's claim that deep networks exhibit a clean two-phase fitting/compression dynamics, and the empirical situation remains disputed. Tishby's framework is invoked here as a unifying objective, not as a settled mechanistic explanation of deep-learning generalization. Treat it as the cleanest articulation we have, not as decided physics.

3. **"Intrinsic dimensionality" is not one number.** The 8–12 dimensions reported for motor cortex, the ~16 dimensions Kaushik finds in weight space, the ~19 dimensions Gong et al. find in ResNet image embeddings, and the ~10 dominant eigenmodes Stringer et al. find in mouse V1 are computed by *different* methods (linear PCA, nonlinear ID estimators, spectral decomposition of weight matrices, cross-validated PCA). The structural-isomorphism claim does not depend on these numbers being identical; it depends on them all being orders of magnitude smaller than ambient dimensionality, which they uniformly are. Quote them as same-order-of-magnitude convergence, not as exact equality.

4. **Stringer et al.'s "high-dimensional" V1 finding initially sounds contradictory.** Their 2019 *Nature* paper title says high-dimensional, and the 1/n eigenspectrum power law does mean information is spread across many dimensions. Steinmetzlab The reconciliation — also addressed in Gauthaman & Ménard's 2024 re-analysis — is that the *dominant* dimensions are few (~10 carry ~30% of variance) biorxiv but the tail is long

and the smoothness constraint (slope α≈1) is what prevents low-dimensional collapse. bioRxiv The cortical manifold is low-effective-dimensional with a long tail, not low-dimensional in the strictest sense. This is consistent with — and arguably exactly what one would predict from — a system optimizing the information-bottleneck objective at the edge of generalization.

5. **The 20-watt figure is an aggregate; the cortical computation fraction is much smaller.** Use Levy & Calvert's 2021 audit (~0.1–0.2 W actual computation, ~3.5 W communication) when precision matters. The 20 W headline is correct as the brain's total metabolic load but conflates computation, communication, and housekeeping in a way that can be misleading in argument.

6. **The Friston free energy principle is a unification claim that not everyone accepts.** It is included here because its mathematical core (variational free energy minimization) genuinely overlaps with the information bottleneck and with predictive coding. Strong-form claims that FEP is a non-falsifiable mathematical principle akin to the principle of least action Wikipedia are controversial; weak-form claims that it captures a common objective across multiple theories of brain function are well-supported.

7. **Hassabis's quotes are accurately transcribed from the public Lex Fridman transcript** (lexfridman.com/demis-hassabis-2-transcript) and from Podscripts of the same episode. The "lower dimensional manifold… true of most of reality" is at approximately the 13:16 mark; the "survival of the stablest" framing at 04:03; "information is primary, more fundamental than energy and matter" Lex Fridman at 06:17. These should be cited by timestamp and source for the Substack version.

8. **The 500-million-year cortex framing rests on the lamprey-pallium homology** established by Suryanarayana, Robertson, Wallén, and Grillner (Nature Ecology & Evolution, 2020) and on broader vertebrate-brain phylogenetic work. The neocortex per se is younger (~200 million years, mammalian); the basal ganglia and the basic pallial plan are the older structures. Use "vertebrate cortex" or "the basic vertebrate forebrain plan" rather than "neocortex" when invoking the longest timescale.

9. **The Kaushik paper is eight months old and the universality finding, while cleanly demonstrated across 1,100+ models, is one paper.** It is being treated here as the AI-side empirical anchor because its result is unusually clean and its method (mode-wise spectral analysis) is reproducible. As with any single paper, replication and extension over the next year will determine whether the ~16-dimensional figure holds across architectures beyond Mistral, ViT, and LLaMA. The structural-isomorphism argument does not require k=16 exactly; it requires the universality and low-dimensionality of the convergence, which Kaushik et al. demonstrate at scale.

10. **Avoid the "we only use 10% of our brain" myth.** It is wrong: imaging shows essentially

all cortical regions are used across the day. The correct statement — that fewer than ~1% of cortical neurons are *substantially active at any given moment* — is the rigorous form Lennie established, and it is the form that connects to sparse coding. This dossier flags the myth only to discharge it; the real arithmetic is far more interesting and is what does the work.