Experiential Measure on the Structure Space of Self-Modeling Systems

Bryan Ehrlich

Abstract

If observer-weight attaches to isomorphism classes of self-modeling dynamical structures rather than to spacetime instantiations, several cosmological pathologies dissolve. This paper formalizes that conditional.

We define a structure space $\mathcal{S}$ of composite Markov processes (B, M) - a body component tracked by a model component - identified up to product-preserving isomorphism. The body/model decomposition is not imposed externally; it is recovered from a factorization condition on the transition kernel. On $\mathcal{S}$, isomorphic systems are a single point by construction - copies collapse. The density functional

$$\rho(p) \;=\; I_p(B;\, M)\,\big(1 - I_p(B;\, M)\,/\,H_p(B)\big)$$

peaks at intermediate self-modeling fidelity and vanishes at both extremes: where there is no self-model and where the self-model is trivially perfect. We demonstrate discriminative power on a toy model, show that unrestricted maximization over partitions fails, and present a conditional proof sketch for Boltzmann brain negligibility via metastability theory.

A companion paper [7] proves that faithful self-modeling on a finite-dimensional spectral order-unit space forces complex quantum mechanics: the state space must carry the structure of a C*-algebra, with the Born rule following via Gleason's theorem. This result sharpens ρ's role. The density selects self-modelers; the algebraic consequences of self-modeling produce quantum mechanics by theorem. When the algebraic structure is pinned to the exceptional Jordan algebra $h_3(\mathbb{O})$ (companion Paper 7), the density resolves uniquely: $\rho_J = \det(X)\,(\mathrm{Tr}(X^2) - \frac{1}{3})$ is the unique lowest-degree F_4-invariant with the self-modeling boundary conditions, producing a falsifiable cubic-vs-quadratic prediction that distinguishes this framework from integrated information theory. The framework does not construct a universal prior ν over structures. The results here - BB suppression, copy identification, density discrimination, and the ρ_J uniqueness theorem - are within-structure claims that do not require ν.

A note on authorship

I am a software engineer, not a mathematician or physicist. AI tools assisted with formal exposition; the ideas and editorial decisions are mine. Feedback on the mathematics is especially welcome.

1 Introduction

1.1 The question

Cosmological measure problems share a common structure: naive counting of observer-instances on spacetime produces paradoxical or empirically wrong answers. Tegmark's Level IV multiverse [23, 24] asks which mathematical structures get weight. Many-worlds branch counting gives uniform probabilities instead of the Born rule. Boltzmann brain arguments show that spacetime-counting makes almost every observer a thermal fluctuation with false memories [5].

This paper proposes an alternative: stop counting instances on spacetime. Instead, place the measure on *isomorphism classes of self-modeling dynamical structures*. A companion result [7] shows that self-modeling is not merely a selection criterion - it is a forcing condition. Faithful self-modeling on a finite-dimensional spectral order-unit space forces a C*-algebraic state space (complex quantum mechanics), with the Born rule following from Gleason's theorem [14]. The density ρ selects self-modelers; the algebra does the rest. This paper constructs the machinery for the selection side and tests it against adversarial cases.

A full measure on structure space would take the form

$$\mathbb{P}(A) \;\propto\; \int_{[x]\in A} \frac{W([x])}{|\mathrm{Aut}(x)|}\, \mathrm{d}\nu([x]) \tag{1}$$

where $\mathcal{S}$ is the structure space, W is the experiential functional, $1/|\mathrm{Aut}|$ is the groupoid weighting, and ν is a reference measure (prior) over structures. This paper constructs $\mathcal{S}$ (Section 2.2), W (Sections 3 and 4), and the quotient (Section 2.4). It does not construct ν. The core results - BB suppression, copy identification, discriminative power - are within-structure claims that don't require it.

1.2 Why self-modeling

We use "self-model" in an operational sense: an internal subsystem M whose state carries information about, and is used to regulate, a set of "body" variables B. This is weaker than "representation" or "consciousness." It means: a subsystem that tracks body variables and whose tracking is recoverable from the system's dynamics.

The motivation comes from cybernetics and control theory. Conant and Ashby's good regulator theorem establishes that every good regulator of a system must be a model of that system [9]. The internal model principle provides the modern formalization: effective regulation requires an internal model that reconstructs the dynamics being regulated [12]. These results have specific technical assumptions. We don't claim they apply universally to all far-from-equilibrium systems. We claim something narrower: candidates for observer-hood in this framework must contain a regulatory subsystem recoverable as a body/model factorization (Section 2.1).

Many systems have such structure. A cell has chemical gradients detecting metabolic state. Animals have rich self-models: hunger tracks metabolic state, pain tracks tissue damage, proprioception tracks limb position [17, 10]. Thermodynamically stable objects - rocks, crystals, dead stars - don't actively regulate anything and receive $\rho = 0$. This is

not a claim that every persisting non-equilibrium structure is conscious. Evolution sharpens self-models (organisms that model themselves better outcompete those that don't), but the basic requirement is operational: active self-maintenance involves self-monitoring.

Formally, a system with a body component B and a model component M that tracks B is a composite Markov process satisfying a factorization condition on its transition kernel (Section 2.1). The factorization detects *tracking structure* - an "observe-then-update" architecture. It doesn't by itself distinguish genuine modeling from passive readout or correlated auxiliary registers. Strengthening the criterion to require bidirectional coupling and a regulation advantage is a natural next step (Section 11.3). The density $\rho = I(B; M)(1 - I/H)$ peaks where self-modeling is rich but incomplete. It vanishes when there is no self-model (left zero) and when the self-model is trivially perfect (right zero). This shape isn't stipulated to peak at humans. It falls out of measuring self-modeling fidelity: the product of "how much does the model know" times "how much doesn't it know" is maximized in the middle.

1.3 Three versions of the same problem

The measure problem appears in three guises.

Tegmark's measure problem. If all consistent mathematical structures exist [23, 24], which ones get more weight? Without a measure over structures, "everything exists" predicts nothing.

The many-worlds counting problem. When a wavefunction branches, naive instance-counting assigns equal probability to all branches. Observation matches the Born rule (probability proportional to amplitude squared), not uniform branch-counting. In this framework, the Born rule is not an additional input: it follows from Gleason's theorem [14] applied to the C*-algebra that self-modeling forces [7].

The Boltzmann brain problem. In an infinite universe or under eternal inflation, thermal fluctuations produce more random brains (momentary configurations with false memories) than evolved ones embedded in functioning ecosystems. Spacetime-counting makes almost every observer a Boltzmann brain [5].

1.4 Axioms and invariances

The framework rests on three postulates and two derived invariances.

Postulates.

P1. Structural experience. What it is like to be a system depends only on its dynamics - transition structure, stationary behavior, self-model fidelity - all of which are preserved under product-preserving isomorphism. Two isomorphic composite processes are one point in the structure space $\mathcal{S}$, not two. See Section 2.3.

P2. Markov level. The composite Markov process (Ω, B, M, Q) is a complete dynamical description at a specified level of coarse-graining. The distribution p_t is the state at this level - not epistemic uncertainty about a finer microstate. The experiential density $\rho(p_t)$ is a property of this level. See Section 4.1.

P3. Factorization witness. We restrict attention to body/model decompositions satisfying the factorization condition $P = P_B \cdot P_M$ (discrete) or the bipartite jump structure (continuous). When such a decomposition exists, it can be recovered by search over product decompositions of the state space. See Sections 2.1 and 3.7.

Derived invariances.

- *Isomorphism invariance.* ρ and μ depend only on the equivalence class $[P] \in \mathcal{S}$, not on the representative. This follows from P1 and the definitions.
- *Time-discretization invariance.* μ is defined as a continuous-time integral (rate matrix Q). The discrete-time sum is a Riemann approximation with fixed physical dt. No dependence on arbitrary time-step choice. See Section D.11.

What is taken as physically real: the rate matrix Q (or kernel P), the product structure (B, M), and the distribution p_t at the Markov level. What is *not* taken as physically real: spatial location, substrate composition, or multiplicity of instantiation.

1.5 Scope and limitations

This paper proposes an axiomatization of observer-weight under a structuralist premise. It delivers definitions, a toy model, and proofs. The principal limitations:

1. **Reference measure (ν).** A full probability measure on $\mathcal{S}$ requires a prior ν over structures (Section 2.5). We don't construct ν. The results in this paper - BB suppression, copy identification, density discrimination - are within-structure claims that don't require it. A universal prior would extend the framework to cross-structure typicality predictions.
2. **Proof status.** The BB negligibility result (Section 7) is presented here as a proof sketch. A complete self-contained proof assembling all seven lemmas with explicit constants is provided in the companion paper [8].
3. **Extraction map.** The framework assumes that composite Markov processes can be identified within larger physical systems. The map from a cosmological system to its composite-process representation is assumed, not constructed (Section 8).
4. **Finite-state and reversible scope.** All definitions and the proof sketch are stated for finite Markov chains with reversible Metropolis-type dynamics. Extensions to continuous state spaces and non-equilibrium dynamics are deferred (Section 11.3).
5. **Empirical breadth.** The toy model consists of seven hand-designed 16-state systems plus a three-state metastability verification. This is sufficient as proof-of-concept. It is not sufficient to establish generality.

2 The structure space

2.1 Composite Markov processes

Notation. We use Ω for state spaces, Q for rate matrices (continuous-time generators), P for discrete-time transition kernels, and $\mathcal{S}$ for the structure space (quotient of composite Markov processes). Ω is a finite set.

We work in the category of finite Markov processes. The formal framework uses continuous-time chains (generator Q, Section 4.1), which eliminates dependence on arbitrary time discretization. The toy model uses discrete-time chains with a fixed physical time step $\mathrm{d}t = 1$ as a special case.

Definition 1 (Composite Markov process)**.** *A* ***composite Markov process*** *is a tuple* (Ω, B, M, P, D) *in the discrete-time case, or* (Ω, B, M, Q, D) *in the continuous-time case, where:*

- $\Omega = B \times M$ *is a finite state space with product structure,*
- B *is the* body *component (the system being modeled),*
- M *is the* self-model *component (the subsystem that represents* B*),*
- P *is a Markov transition kernel on* Ω*,*
- D *is the* decomposition data*: the product structure* $B \times M$ *as a witness for the factorization condition below.*

The transition kernel P *admits the* ***factorization****:*

$$P\big((b', m') \mid (b, m)\big) \;=\; P_B(b' \mid b,\, m) \;\cdot\; P_M(m' \mid b',\, m). \tag{2}$$

This encodes a system with an internal self-model: M *is a driven subsystem that updates based on* B*'s current state. The factorization is the "observe-then-update" structure - M sees what B just did and adjusts accordingly. The one-step kernel* P *is the composition of two substep kernels: first* P_B *(body transitions), then* P_M *(model observes and updates). Both* B *and* M *may change in a single time step because the step encompasses two sequential substeps, not a single simultaneous transition.*

The structural constraint is a conditional independence condition: the model's next state is independent of the body's previous *state given the body's new state and the model's current state:*

$$m' \perp b \mid (b',\, m) \qquad \forall\, b,\, b',\, m,\, m'. \tag{3}$$

Any Markov kernel admits a chain-rule factorization $P(b', m' \mid b, m) = P(b' \mid b, m)\, P(m' \mid b', b, m)$*. The nontrivial content of Eq. 2 is that the second factor drops the dependence on* b*.*

The continuous-time analogue uses a bipartite jump structure on the generator Q (Appendix D.1).

Definition 2 (Isomorphism)**.** *Two composite Markov processes* $(\Omega_1, B_1, M_1, P_1, D_1)$ *and* $(\Omega_2, B_2, M_2, P_2, D_2)$ *are* ***isomorphic*** *if there exists a bijection* $\varphi\colon \Omega_1 \to \Omega_2$ *that:*

1. *preserves the product structure:* $\varphi = \varphi_B \times \varphi_M$ *for bijections* $\varphi_B \colon B_1 \to B_2$ *and* $\varphi_M \colon M_1 \to M_2$*;*

2. *preserves the kernel:* $P_2(\varphi(s'), \varphi(s)) = P_1(s', s)$ *for all* s, s'*.*

This is finer than unrestricted state relabeling. An arbitrary bijection could scramble body and model components, mixing the two subsystems and destroying the factorization structure. The product-preserving constraint prevents this (see Section 3.7 for why this matters).

2.2 The structure space

Definition 3 (Structure space)**.** *The* ***structure space*** $\mathcal{S}$ *is the quotient*

$$\mathcal{S} \;=\; \{\textit{composite Markov processes}\} \,/\, \sim$$

where $\sim$ *is isomorphism as defined above.*

The topology of $\mathcal{S}$ (orbifold structure, stratification) is discussed in Appendix D.2.

A Markov kernel on N states that admits no nontrivial product decomposition satisfying the factorization condition has no self-model. Such systems are included in $\mathcal{S}$ via the trivial decomposition $M = \{*\}$ (single-element model component), which gives $I(B;\, M) = 0$ and therefore $\rho = 0$. This preserves the $\Omega = B \times M$ form of Definition 1 while correctly assigning zero experiential weight.

2.3 Multiplicity as gauge (postulate)

A central modeling choice is that multiplicity of isomorphic realizations is treated as non-physical for the purpose of observer-weighting: the measure is defined on equivalence classes in $\mathcal{S}$ rather than on spacetime instantiations. Equivalently, "copy number" is treated as a gauge degree of freedom that is removed by quotienting.

The motivation (P1): what it is like to be a system depends on its dynamics – transition structure, stationary behavior, self-model fidelity – all preserved under isomorphism. Two instances of the same composite Markov process on different substrates have identical dynamical structure. Weighting them separately would be weighting the same structure twice because it happens to occur at different spacetime addresses. On $\mathcal{S}$, they are the same point, by definition of the quotient.

Quotienting by symmetry is standard in physics: the Gibbs factor $1/N!$ corrects for overcounting identical microstates; Bose–Einstein and Fermi–Dirac statistics arise from quotienting by particle permutation symmetry; Feynman diagram symmetry factors are groupoid cardinality [2]. But those corrections adjust the entropy while keeping thermodynamic quantities extensive in N. Our premise is stronger: isomorphic composite processes are *the same point* in $\mathcal{S}$, and parallel instantiations do not additively increase the experiential measure.

This is a postulate about what the measure should depend on, not a derivation. Readers who hold that distinct spacetime instantiations of the same dynamics necessarily contribute additively should treat the remainder of this paper as exploring an alternative convention.

2.4 Groupoid cardinality

When counting objects up to isomorphism, the correct mathematical tool is groupoid cardinality [2]. The objects of the relevant groupoid are *isomorphism classes* of composite processes, not individual instantiations across spacetime. Morphisms are automorphisms within each class. For a groupoid $\mathcal{G}$:

$$|\mathcal{G}| \;=\; \sum_{[x]} \frac{1}{|\mathrm{Aut}(x)|} \tag{4}$$

where $\mathrm{Aut}(x)$ is the automorphism group of x. A complex, asymmetric system (like a brain) has $\mathrm{Aut} = \{\mathrm{id}\}$ and counts as 1. A highly symmetric system has large $|\mathrm{Aut}|$, and its contribution is suppressed.

The $1/|\mathrm{Aut}|$ weighting follows from G-invariant measures on the kernel space (Appendix D.3).

2.5 What we do and do not construct

We construct two things cleanly:

1. A **quotienting principle**: isomorphic composite Markov processes are identified, collapsing copies.
2. An **intra-system experiential functional**: given a composite Markov process, ρ and the trajectory integral are computable scalars (Sections 3 and 4).

We do **not** construct a full probability measure on $\mathcal{S}$. That would require a **reference measure** (or prior) ν on $\mathcal{S}$ - e.g., algorithmic probability, maximum-entropy over kernels, or Solomonoff-type enumeration. A full measure would take the form

$$\mathbb{P}(A) \;\propto\; \int_{[x]\in A} \frac{W([x])}{|\mathrm{Aut}(x)|}\,\mathrm{d}\nu([x]) \tag{5}$$

where $W([x])$ is the expected experiential functional (Section 4) and ν is the reference measure. We define W and the quotient. We do not define ν.

Even without ν, the framework is not vacuous. The quotient collapses copies, eliminating one divergence class. The density W discriminates self-modeling systems from non-self-modeling ones. Any candidate ν inherits these constraints.

3 The experiential density ρ

3.1 Definition

Definition 4 (Experiential density)**.** *The* ***experiential density*** *is a functional of the joint distribution over (B, M) at any given time. Given a distribution p over $\Omega = B \times M$:*

$$\rho(p) \;=\; I_p(B;\, M)\left(1 - \frac{I_p(B;\, M)}{H_p(B)}\right) \tag{6}$$

where $I_p(B;\, M)$ is the mutual information between body and self-model under p, and $H_p(B)$ is the Shannon entropy of the body's marginal distribution under p.

Equivalently, $\rho(p) = I_p(B;\, M) \cdot H_p(B \mid M)/H_p(B)$. The first factor measures represented information: how much does the self-model know about the body? The second factor measures remaining uncertainty: how much does the self-model *not* know? The product peaks when both are substantial.

At stationarity ($p = \pi$), we write $\rho(\pi)$ or simply ρ when the stationary distribution is understood. For a non-stationary trajectory, $\rho_t := \rho(p_t)$ where p_t is the distribution over (B, M) at time t.

3.2 Properties

For any distribution p over $\Omega = B \times M$:

- $\rho(p) = 0$ when $I_p(B;\, M) = 0$: no self-model. The model carries no information about the body.
- $\rho(p) = 0$ when $I_p(B;\, M) = H_p(B)$: perfect self-model. The model compresses the body losslessly.
- $\rho(p)$ peaks at $I_p(B;\, M) = H_p(B)/2$: rich but incomplete self-model.
- $\rho(p)$ is bounded: $0 \leq \rho(p) \leq H_p(B)/4$.
- ρ is invariant under product-preserving isomorphism: if $\varphi = \varphi_B \times \varphi_M$ conjugates P_1 to P_2, then $\rho(\pi_1) = \rho(\pi_2)$.
- Edge case: when $H_p(B) = 0$ (degenerate body), define $\rho(p) = 0$ by continuity.
- ρ scales with accessible body variability: a system trapped in a small subset of body states has low $H_p(B)$ and therefore low ρ, even if its self-modeling is rich relative to that subset. This is a substantive design choice – experiential weight requires body variability, not just modeling fidelity.

The normalization by $H(B)$ creates a coarse-graining vulnerability (state padding) discussed in Appendix D.4.

3.3 The right zero: design choice, not metaphysics

The right zero ($\rho = 0$ at $I = H$) is the most philosophically loaded feature of the density. The operational justification: ρ should not reward trivially predictable systems. A crystal where $M = B$ (perfect copy) has $I(B;\, M) = H(B)$, but it's doing nothing interesting - it's a lookup table, not a model. A limit cycle with deterministic self-tracking is the same. The right zero ensures these systems score zero, just as a system with no self-model scores zero. What ρ rewards is the middle: systems that know something about themselves but not everything, where the self-model is doing real computational work.

This is a design choice in the effective-complexity / LMC tradition (see 3.4), not a theorem about consciousness. Other density functions with the same qualitative shape would serve the same role in the BB suppression theorem, provided ρ satisfies the boundary-vanishing condition.

The qualitative properties that matter are formalized as an admissible density family (Appendix D.5).

3.4 This form is not arbitrary

The density function belongs to a recognized family:

Effective complexity [15]. The algorithmic information content of regularities, which peaks between pure order and pure randomness. The $I(1 - I/H)$ form is the information-theoretic realization of effective complexity.

LMC statistical complexity [19]. The entropy-times-disequilibrium form $H \cdot D$ that peaks at intermediate complexity. Our density has the same qualitative structure.

3.5 Why self-modeling, not integration

Tononi's Integrated Information Theory (IIT) uses Φ (integrated information) as a measure of consciousness [25]. Φ asks "how much information does the system integrate across its parts?" without asking "information about what?" This creates the Aaronson problem: simple XOR gate grids score high on Φ because they integrate lots of information about their inputs, despite having no self-referential structure.

Our density measures something more specific: mutual information between a system and its model *of itself*. $I(B;\, M)$ is low for XOR grids (no meaningful self-model exists) and $I(B;\, M) \approx H$ for perfectly regular structures. Either way, $\rho \approx 0$ for the simple repetitive structures that IIT flags.

3.6 Representational vs. predictive information

The framework admits two co-equal density functionals:

$$\rho(p) = I_p(B_t;\, M_t)\left(1 - I_p(B_t;\, M_t)/H_p(B_t)\right) \qquad \text{(representational)} \qquad (7)$$

$$\rho_{\text{pred}}(p) = I(M_t;\, B_{t+1})\left(1 - I(M_t;\, B_{t+1})/H(B_{t+1})\right) \qquad \text{(predictive)} \qquad (8)$$

The representational density measures how well the model represents the body's current state. The predictive density measures how well the model predicts the body's *future* state. We compute ρ_{pred} for all seven systems in the toy model. The joint distribution over (M_t, B_{t+1}) is

$$P(M_t = m,\, B_{t+1} = b') \;=\; \sum_b \pi(b, m)\, P_B(b' \mid b,\, m),$$

from which $I(M_t;\, B_{t+1})$ is computed directly. Results (Table 1):

The separation under ρ_{pred} is dramatic: >600× ratio between Observer and BB+self-model (0.289 vs. 0.001), compared to 1.5× under ρ (0.347 vs. 0.239). The BB's random body has no temporal structure for the model to predict - $I(M_t;\, B_{t+1})$ collapses to near zero.

System	ρ (representational)	ρ_{pred} (predictive)
Observer	0.347	0.289
BB + self-model	0.239	0.001
Crystal	0.055	0.174
All others	≈ 0	≈ 0

Table 1: Representational vs. predictive density for the seven toy-model systems.

Two independent suppression channels.

- ρ **+ metastability (Theorem A):** BB suppression via trajectory duration. The chain spends exponentially more time in the stable basin than in BB excursions. This works even when instantaneous ρ is non-trivial.
- ρ_{pred}**:** BB suppression via density alone. Random bodies have no temporal structure, so predictive MI collapses - $\rho_{\text{pred}} \approx 0$ for BB+self-model without invoking metastability.

A critic of the metastability argument has a density-only fallback (ρ_{pred}). A critic of predictive MI has a duration-based fallback (ρ + Theorem A). The BB+self-model case is handled by either route.

Predictive MI is a BB filter, not a triviality filter. The crystal's $\rho_{\text{pred}} = 0.174$ (Table 1) shows that ρ_{pred} rewards highly regular systems: a near-deterministic cycle is maximally predictable, so predictive MI is high. The representational right zero ($\rho = 0$ at $I = H$) suppresses trivially predictable systems; ρ_{pred} does not. No single simple functional captures all desiderata. The representational and predictive densities address complementary failure modes. A combined criterion - requiring nontrivial representational richness *and* nontrivial prediction - is the natural strengthening, and would suppress BBs and trivially regular systems simultaneously.

3.7 The subsystem decomposition

An objection: who decides the body/model partition? If ρ depends on a choice of labeling, it's not a property of the system. For ρ to be well-defined, the (B, M) decomposition must be identifiable from the transition matrix alone.

It is. The factorization condition (Eq. 2) is checkable from P alone. Given a transition matrix on N states, enumerate all product decompositions $N = |B| \times |M|$, test which ones satisfy the factorization, and among valid factorizations compute ρ. Systems with no valid factorization get $\rho = 0$.

The obvious alternative - maximizing ρ over *all* possible partitions regardless of dynamics - fails. We tested this (Section 5.8). A crystal with near-perfect self-tracking achieves $\rho \approx 0$ under its dynamically realized partition ($I/H \approx 1$, right zero). But an external analyst can scramble the coordinates and find a partition where the same correlations look like partial self-modeling ($I/H \approx 0.5$, peak). The scrambled partition exploits raw correlation without respecting the causal structure.

Non-uniqueness. Multiple valid factorizations of the same kernel may exist. We take ρ as the maximum over all valid factorizations. This is well-defined (finitely many product decompositions) and invariant under isomorphism.

Details of approximate factorization, selection principles, and stability conjectures are in Appendix D.6.

4 The trajectory functional

The measure must integrate over trajectories, not snapshots.

4.1 Definition (continuous-time)

Given a composite process (Ω, B, M, Q, D) where Q is the rate matrix of a continuous-time Markov chain, the distribution evolves as $\mathrm{d}p/\mathrm{d}t = pQ$, with solution $p_t = p_0\, e^{Qt}$.

Definition 5 (Experiential functional)**.** *The* ***experiential functional*** *over a time interval* $[0, T]$ *is*

$$\mu([0, T]) \;=\; \int_0^T \rho(p_t)\, \mathrm{d}t \tag{9}$$

where $\rho(p_t) = I_{p_t}(B;\, M)\,(1 - I_{p_t}(B;\, M)/H_{p_t}(B))$*. The* ***stationary rate*** *is* $r(Q) = \rho(\pi)$ *where* $\pi Q = 0$*.*

The functional μ is well-defined without any path measure or expectation operator: it is an integral of information-theoretic quantities computed from the marginal distribution at each instant. The integrand $\rho(p_t)$ is continuous in t (since $p_t = p_0\, e^{Qt}$ is smooth and mutual information is continuous in the distribution), so the integral exists as a standard Riemann integral. The units are information $\times$ time (nat-seconds when using natural logarithms).

The Markov level axiom (P2), the distinction between distributional and path-level functionals, and time-rate sensitivity are detailed in Appendix D.7.

4.2 Normalization and comparison

At stationarity, $\mu([0, T])$ grows linearly with T. For comparing different systems, we use either matched-horizon ratios $\mu([0, T]; Q_1)/\mu([0, T]; Q_2)$ for the same T, or stationary rates $r(Q)$ directly.

For the BB suppression argument (Section 7), the relevant quantity is μ accumulated during different phases of a non-stationary trajectory: residence in the stable basin vs. BB excursions.

4.3 Why trajectories, not snapshots

The trajectory formulation is what distinguishes our approach from static observer-counting. A Boltzmann brain that momentarily achieves high ρ contributes for a brief spike duration. A stable observer in a deep basin contributes $\rho(\pi)$ for the full basin residence time. Even when peak ρ is comparable (Section 5.5), the integrated functionals differ by the ratio of their durations, which is exponential (Section 7).

5 Toy model: composite-state self-modeling

5.1 Setup

To test the density function's discriminative power, we constructed seven 16-state Markov chains with composite state space $\Omega = B \times M$ where $B = \{0, 1, 2, 3\}$ and $M = \{0, 1, 2, 3\}$. Each system has a 16×16 transition matrix P. States are indexed as $(b, m) \mapsto 4b + m$, giving state indices 0 through 15. For each system, we compute the stationary distribution π, the marginals, $I_\pi(B;\, M)$, $H_\pi(B)$, and $\rho(\pi)$.

All systems use the observe-then-update (no-lag) formulation:

$$P\big((b', m') \mid (b, m)\big) \;=\; P_B(b' \mid b) \;\cdot\; P_M(m' \mid b',\, m). \tag{10}$$

The model subsystem sees the *new* body state b' and updates accordingly. This corresponds to a two-stage update within each time step (body transitions, then model observes and updates). An earlier version with one-step lag introduced a Data Processing Inequality artifact that capped $I(B;\, M)$ below $H(B)/2$ for metastable body dynamics.

5.2 Body and model dynamics

Three body kernels are used: *metastable* (two basins with rare cross-basin transitions), *cyclic* (near-deterministic $0 \to 1 \to 2 \to 3 \to 0$), and *random* (near-uniform). For tracking models, the model update given the new body state b' and old model state m is a mixture: snap to b' with probability α (tracking accuracy), persist at m with probability β, and uniform noise with probability $\gamma = 0.02$. The parameter α controls self-model fidelity: $\alpha \to 0$ means no tracking, $\alpha \approx 0.5$ means imperfect tracking (peak ρ), and $\alpha \to 1$ means near-perfect copying (right zero). For independent models, M runs its own dynamics unrelated to B. Full transition matrices and a worked example are in Appendix D.8.

5.3 The seven systems

1. **Self-modeling observer.** Metastable body (Eq. 15), model tracks body with $\alpha = 0.5$. The model updates based on what the body does, but imperfectly.

2. **Thermostat.** Metastable body, model runs an independent cycle unrelated to body. The model tracks *something*, but not itself.

3. **Crystal (XOR analog).** Cyclic body (Eq. 16), model tracks with $\alpha = 0.99$, $\gamma = 0.005$. The system "knows itself" almost perfectly.

4. **Boltzmann brain.** Random body, random independent model. Both components are near-uniform.

5. **Smart thermostat.** Cyclic body (highly predictable), independent model cycle. Body is predictable, but the model doesn't track it. This system kills any definition based on $I(X_t;\, X_{t+1})$ (self-predictability rather than self-reference).

6. **BB with self-model.** Random body, model tracks body with $\alpha = 0.5$. The body is unstructured, but the model tries to follow it.

7. **Buffeted (adversarial test).** Metastable body; the body *scrambles* the model via Klein four-group permutations $\{e, (01)(23), (02)(13), (03)(12)\}$. Each body state applies a different permutation to M. There is coupling (M's transitions depend on B), but the coupling destroys information rather than transmitting it. The Klein four-group acts transitively on $\{0, 1, 2, 3\}$, sending any M distribution toward uniform when applied by a randomly-selected body state.

5.4 Pre-registered predictions and results

Predictions were written before running the model. Table 2 shows the predictions and Table 3 shows the results.

#	System	Predicted I/H	Predicted ρ	Rationale
1	Observer	0.3–0.7	near peak	Near peak of parabola
2	Thermostat	0	0	Product distribution
3	Crystal	≈ 1	≈ 0	Right zero (perfect model)
4	BB	≈ 0	≈ 0	Left zero (nothing to model)
5	Smart thermostat	0	0	Independence
6	BB + self-model	≈ 0	≈ 0	Nothing structured to track
7	Buffeted	≈ 0	≈ 0	Coupling destroys information

Table 2: Pre-registered predictions, written before computing results.

System	$H(B)$	$I(B; M)$	I/H	ρ
Observer ($\alpha = 0.5$)	1.386	0.705	0.508	0.347
Thermostat (M indep.)	1.386	0.000	0.000	0.000
Crystal ($M \approx B$)	1.386	1.329	0.959	0.055
BB (both random)	1.386	0.000	0.000	0.000
Smart thermostat	1.386	0.000	0.000	0.000
BB + self-model	1.386	0.306	0.221	0.239
Buffeted	1.386	0.000	0.000	0.000

Table 3: Toy model results. Eight of nine pre-registered predictions confirmed.

The observer sits at $I/H = 0.508$, essentially the theoretical peak. The crystal is at the right zero ($I/H = 0.959$, $\rho = 0.055$). The BB, thermostat, smart thermostat, and buffeted system all have $\rho \approx 0$, for the correct reasons. The BB+self-model result ($\rho = 0.239$, $I/H = 0.221$) is the one "failure" - see Section 5.5.

Figure 1 shows the results: bar chart of ρ across all seven systems, positions on the ρ parabola, and the parametric sweep tracing the full curve.

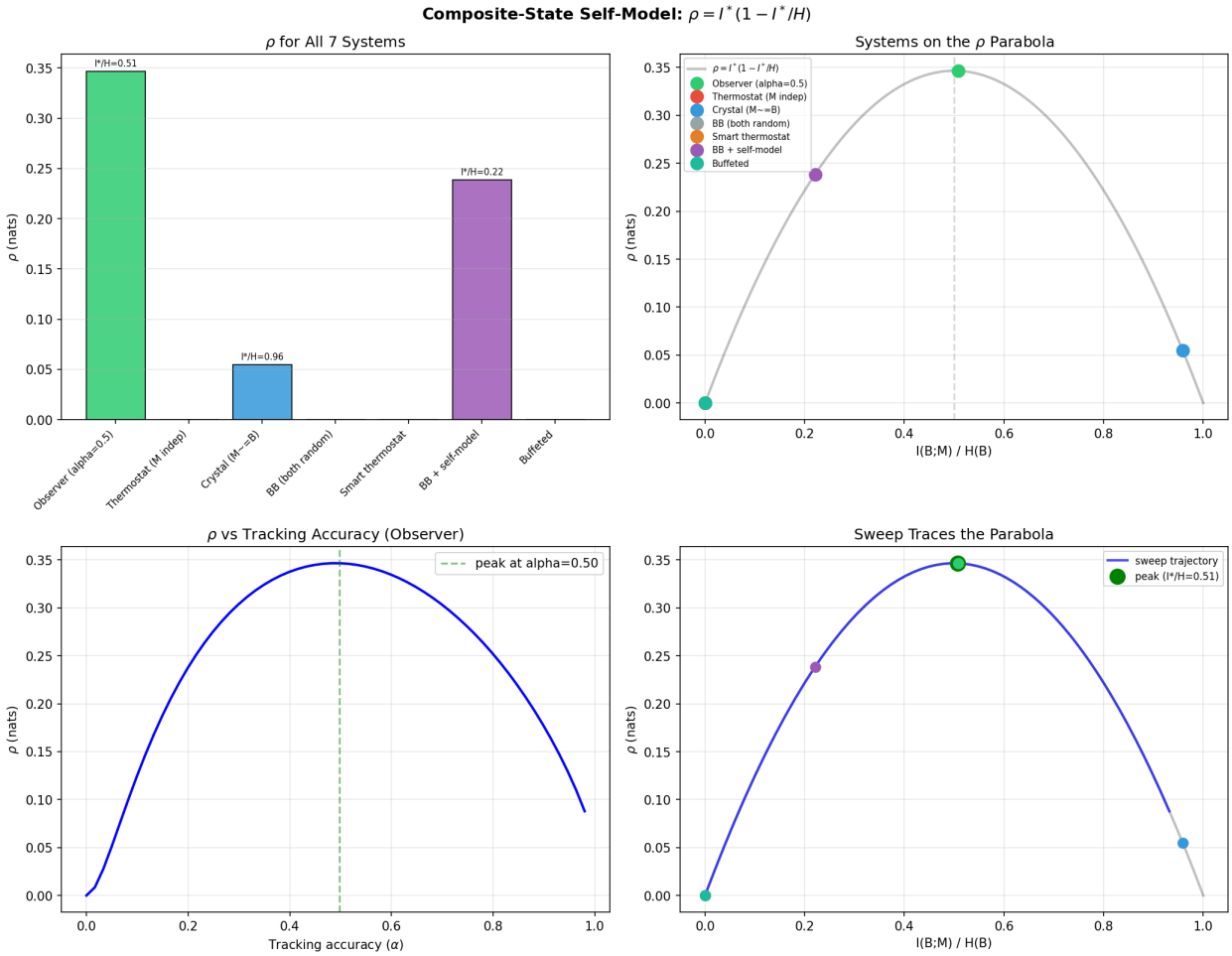

Figure 1: Composite-state toy model results. *Top left:* ρ for all seven systems. *Top right:* Systems plotted on the $\rho = I(1 - I/H)$ parabola. *Bottom left:* ρ vs. tracking accuracy α for the observer, tracing the full parabola from left zero ($\alpha = 0$) through peak ($\alpha \approx 0.5$) to right zero ($\alpha \to 1$). *Bottom right:* The parametric sweep traced on the parabola curve.

5.5 The BB + self-model result

The one "failed" prediction: we predicted $\rho < 0.1$ for the BB with a self-model, but got $\rho = 0.239$ - non-trivial, though lower than the observer's 0.347. This is actually correct, and the prediction was wrong.

A self-modeling system with random body dynamics *does* achieve non-trivial $I(B; M)$ because the model observes and partially tracks the body's current state. The model can't predict the body's future (it's random), but it can represent the body's present. With a near-uniform body, the tracking accuracy yields $I/H = 0.221$ - lower than the observer's 0.508 because the random body provides less structured information, but well above zero.

The formalism doesn't suppress BBs by assigning them zero instantaneous density. It suppresses them via the trajectory integral. A BB-with-self-model achieves non-trivial *instantaneous* ρ (0.239 vs. the observer's 0.347), and its *trajectory* is a spike: ρ rises from ≈ 0 as

the fluctuation assembles, peaks briefly, and returns to ≈ 0 as it dissipates. The integrated measure is $\rho \times$ duration, and the duration is vanishingly short compared to a stable basin's residence time. (Alternatively, the predictive density ρ_{pred} suppresses BB+self-model at the density level alone - see Section 3.6.)

5.6 Parametric sweep

Sweeping the observer's tracking accuracy from $\alpha = 0$ (no tracking) to $\alpha = 0.98$ (near-perfect tracking) traces the full parabola:

- $\alpha = 0$: $I(B; M) = 0$, $\rho = 0$ (left zero).
- $\alpha \approx 0.5$: $I(B; M) \approx H/2$, $\rho \approx H/4$ (peak).
- $\alpha \to 1$: $I(B; M) \to H$, $\rho \to 0$ (right zero).

Peak ρ occurs at $\alpha = 0.498$, with $I/H = 0.507$ - essentially the theoretical maximum at $I = H/2$. The sweep demonstrates that the density function behaves as designed.

5.7 The buffeted adversarial test

The buffeted system is the hardest test. Body and model are coupled (M's transitions depend on B), but the coupling is information-destroying: each body state applies a Klein four-group permutation to M, scrambling M's state rather than informing it.

Result: $I(B; M) = 0.000$. The coupling creates no stationary correlation. Information is destroyed, not transmitted. This confirms that $I(B; M)$ alone suffices to distinguish "tracks" from "buffeted" - we don't need directional measures like transfer entropy for this class of systems.

5.8 Why max-over-partitions fails

An obvious fix to the subsystem decomposition problem: define ρ as the maximum over *all* possible product decompositions of the state space, not just the dynamically realized one. We tested this by evaluating ρ under 200,000 random permutations of the 16-state space, plus hill-climbing optimization from 20 random starts.

For systems without internal structure (thermostat, BB, smart thermostat, buffeted), the max stays at zero regardless of partition. For the observer and BB+self-model, the natural partition is near-optimal (99.8th and 98.7th percentile respectively).

The crystal breaks it. Under its natural partition, the crystal has $I/H = 0.96$ and $\rho = 0.055$ (the right zero). But a scrambled partition redistributes the crystal's strong internal correlations so they look like partial self-modeling: $I/H \approx 0.5$, $\rho = 0.346$. The max-over-partitions crystal scores identically to the observer (Figure 2).

This is why the factorization condition (Section 3.7) is necessary. Maximizing ρ over arbitrary decompositions re-opens the Aaronson failure mode.

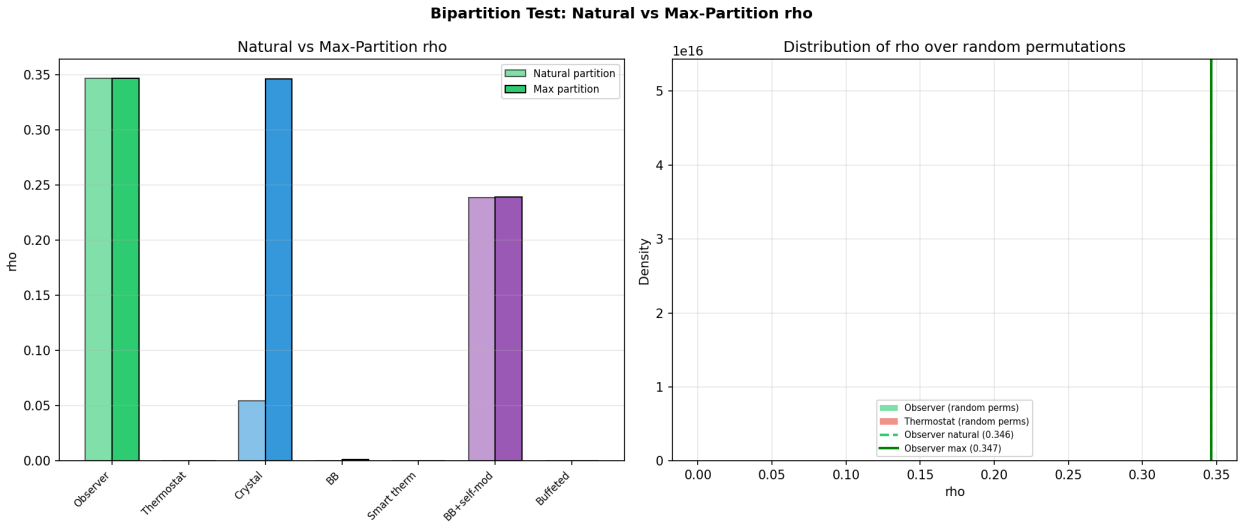

Figure 2: Bipartition test: natural partition vs. max-over-all-partitions. *Left:* The crystal under max-partition scores as high as the observer ($\rho \approx 0.346$), defeating the right-zero suppression. *Right:* Distribution of ρ over random permutations for the observer, showing the natural partition is near-optimal.

6 Copies collapse

Definition 6 (Copy collapse)**.** *Isomorphic composite Markov processes (Definition 2) are identified - they occupy the same point in $\mathcal{S}$. Multiple instantiations of the same system in different spatial locations do not increase the experiential measure.*

This is definitional. It is the content of working on the quotient space.

Proposition 7. *ρ and μ are class functions on $\mathcal{S}$: they depend only on the equivalence class $[P]$, not on the representative.*

Proof. Isomorphism preserves the kernel, stationary distribution, marginals, and product structure, so $I(B; M)$, $H(B)$, and ρ are invariant. The trajectory integral inherits invariance from the pointwise density. □

7 BB negligibility (conditional proof sketch)

7.1 The corrected theorem statement

An earlier version of this work stated the theorem as a $T \to \infty$ limit. This is wrong. By the ergodic theorem, as $T \to \infty$ for a fixed system, the ratio of BB measure to stable measure converges to a *fixed positive constant*. The ergodic theorem doesn't suppress BBs - it enshrines them at their stationary-measure weight.

The correct asymptotic parameter is $\varepsilon \to 0$ (noise/temperature), not $T \to \infty$.

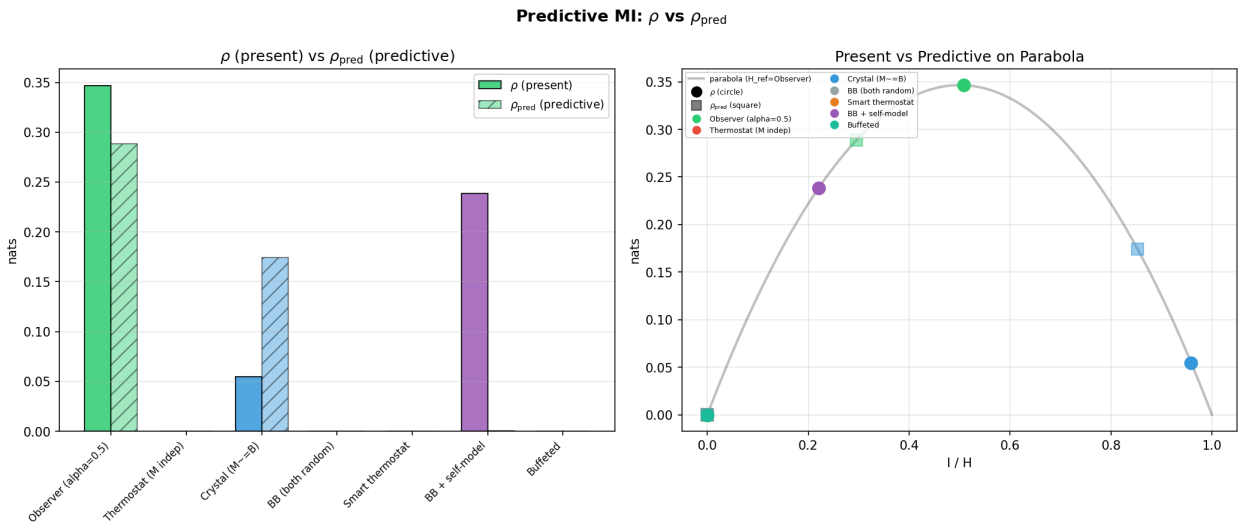

Figure 3: Representational (ρ) vs. predictive (ρ_{pred}) density. *Left:* Bar chart comparison. The BB+self-model separation under ρ_{pred} is >600×, compared to 1.5× under ρ. *Right:* Both densities plotted on the parabola; circles are ρ, squares are ρ_{pred}.

7.2 The claim

The intuition is simple. Model the state space as an energy landscape with deep and shallow wells. Stable observers occupy deep wells – they require stellar environments, complex chemistry, large entropy barriers to maintain. BBs are transient thermal fluctuations sitting in shallow wells. A system with noise intensity ε spends time $\sim \exp(\Delta/\varepsilon)$ in a well of depth Δ before escaping. When the stable well is deeper ($\Delta_s > \Delta_b$), the ratio of time spent in the BB well to the stable well decays exponentially:

$$\frac{\mu_{\text{BB}}}{\mu_{\text{stable}}} \sim \exp\left(-\frac{\Delta_s - \Delta_b}{\varepsilon}\right) \to 0 \quad \text{as } \varepsilon \to 0. \tag{11}$$

The suppression is exponential in the depth gap and the inverse noise. Even when BB states achieve comparable instantaneous ρ, the integrated measure over trajectories is dominated by the stable basin's residence time.

This is a standard consequence of Freidlin–Wentzell metastability theory [13] applied to the trajectory functional. A formal theorem statement, lemma dependency graph, and proof sketch using BEGK potential theory [4] are in Appendix D.10.

7.3 Numerical verification

We verify on a three-state chain ($s \leftrightarrow u \leftrightarrow b$) with Freidlin–Wentzell transition rates, where s is the stable state (deep well, $\Delta_s = 3$), b is the BB state (shallow well, $\Delta_b = 1$), and u is the transition state. The experiential measure ratio has the closed form

$$\frac{\mu_{\text{BB}}}{\mu_{\text{stable}}} = \frac{\rho_b}{\rho_s} \cdot \frac{1-p}{p} \cdot \exp\left(-\frac{\Delta_s - \Delta_b}{\varepsilon}\right). \tag{12}$$

Three suppression mechanisms are visible: (1) metastability $\exp(-(\Delta_s - \Delta_b)/\varepsilon)$ (exponential, does all the work); (2) density ratio $\rho_b/\rho_s < 1$ (multiplicative); (3) branching prefactor $(1-p)/p$ (order-one). Simulated trajectories confirm the analytical formula with $< 1\%$ error, and the ratio decays exponentially. At $\varepsilon = 0.1$, $\mu_{\text{BB}}/\mu_{\text{stable}} \approx 5 \times 10^{-10}$ (Figure 4). Full construction details in Appendix D.9.

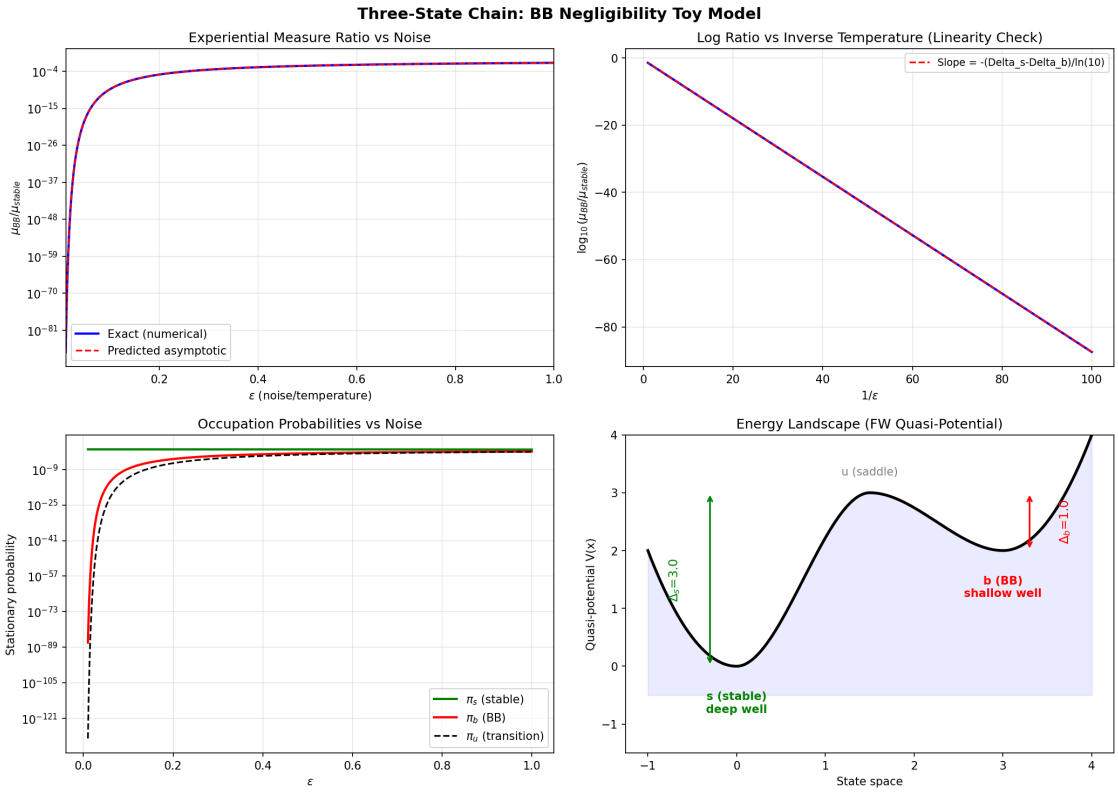

Figure 4: Three-state chain verification. *Top left:* $\mu_{\text{BB}}/\mu_{\text{stable}}$ vs. noise ε, showing exponential decay matching the analytical prediction. *Top right:* $\log_{10}$ of the ratio vs. $1/\varepsilon$, confirming linearity. *Bottom left:* Occupation probabilities vs. ε. *Bottom right:* The energy landscape with deep well ($\Delta_s = 3$) and shallow well ($\Delta_b = 1$).

7.4 The critical assumption

$\Delta_s > \Delta_b$ is an assumption, not a consequence of the framework. The theorem says: "given that stable observers sit in deeper thermodynamic wells than BB fluctuations, the experiential measure of BBs is negligible." The gap is a physics input. In many cosmological discussions this inequality is treated as plausible – brains require stellar environments with large entropy barriers; BBs are transient thermal fluctuations – but establishing it rigorously is model-dependent and outside scope.

For related approaches to the BB problem via observer memory and thermodynamic consistency, see Wolpert & Kipper [29].

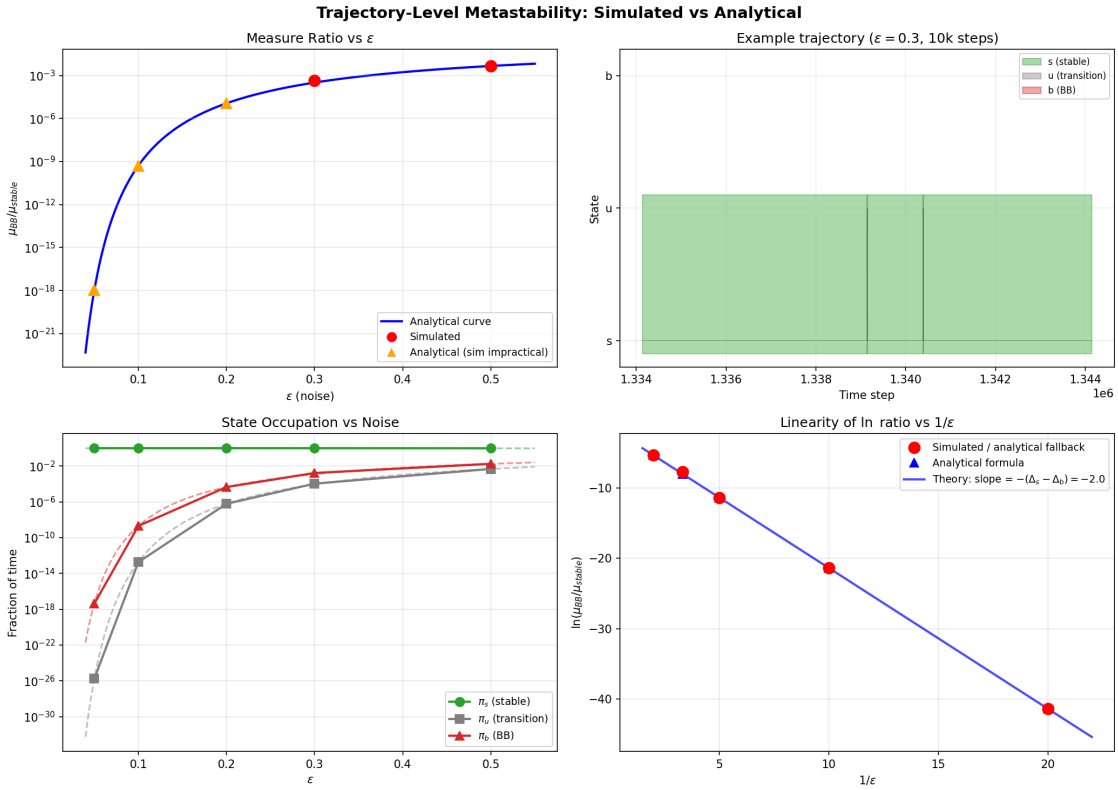

Figure 5: Trajectory-level verification. *Top left:* Simulated measure ratio vs. analytical prediction. *Top right:* Example trajectory at $\varepsilon = 0.3$ showing long residence in the stable state with rare BB excursions. *Bottom left:* State occupation fractions vs. noise. *Bottom right:* Linearity of $\ln(\mu_{\text{BB}}/\mu_{\text{stable}})$ vs. $1/\varepsilon$ with theoretical slope -2.0.

8 Perfect duplicate identification

Two framings of "Boltzmann brain." The framework uses two distinct framings:

- **Standalone composite processes in $\mathcal{S}$.** A BB is a composite Markov process extracted from some larger physical system. Its kernel, decomposition, and stationary distribution define its point in $\mathcal{S}$.
- **Excursions within a single metastable chain.** In Theorem A, "BBs" are transient visits to a shallow metastable region within one large chain.

The extraction map - how to identify candidate composite processes within a larger system - is assumed, not constructed.

Perfect duplicates. A Boltzmann brain that is an exact physical duplicate of a normal brain, viewed as a standalone composite process, has the same transition kernel and therefore the same ρ. The density mechanism doesn't suppress it. The metastability mechanism doesn't either.

Under the structural-experience postulate (P1), a perfect duplicate is isomorphic to the original. On $\mathcal{S}$, they are the same point. This is definitional, not derived - it's what P1 says.

The BBs that survive as distinct points on $\mathcal{S}$ are *imperfect* duplicates: partial fluctuations, degraded copies, brains-in-a-moment with no causal history. For these, ρ is low (incomplete self-model) and duration is vanishing (they dissipate immediately). Both the density and metastability mechanisms suppress them.

Whether collapsing perfect duplicates is a feature or a bug depends entirely on accepting P1. A reader who holds that numerically distinct spacetime instantiations of the same dynamics must contribute additively will reject this move. The rest of the paper still applies to imperfect BBs.

9 Comparison with existing work

9.1 Müller (2020) - closest predecessor

Müller's "Law without Law" derives physics from algorithmic probability of observer states, published in *Quantum* [20]. Same paradigm: observer-centric, structural measure, addresses BBs.

The key difference: Müller's density uses Kolmogorov complexity, which is uncomputable. Ours uses mutual information, which is computable and continuous.

The shared gap: both frameworks provide a weighting functional but leave the reference measure implicit. Neither constructs a full probability measure on structure space. Müller (2026) extends this program [21], strengthening the case that observer-state structure is the correct locus for the measure. Our continuous-time trajectory functional is complementary.

9.2 Tononi's IIT

IIT's Φ measures integrated information across a system's parts [25]. Our density measures mutual information between a system and its *self-model*. The Aaronson problem (XOR grids scoring high on Φ) doesn't arise because XOR grids have no self-model. See Section 5.8.

9.3 Friston's Free Energy Principle

The FEP says persistent Markov-blanketed systems minimize variational free energy, which requires an implicit internal generative model. If this connection holds precisely, the FEP provides a dynamical account of *why* systems self-model; our framework provides a measure-theoretic account of *how much* experiential weight that self-modeling carries. The relationship is suggestive, not formal - formalizing it would require mapping FEP's Markov blanket structure onto our (B, M) factorization, which is not attempted here.

9.4 Mathematical foundations

Baez & Dolan [2] provide the groupoid cardinality framework. Pinto & Rand [22] construct moduli-type parameterizations for hyperbolic dynamical systems. Gell-Mann & Lloyd [15]

ground the density function in established information theory. Ackerman, Freer & Patel [1] develop rigorous measures on spaces of countable structures under permutation invariance - the closest existing mathematics to "measure on structure-space up to isomorphism." Weinstein [28] identifies the technical burden for continuous settings: invariance under equivalence requires extra geometric data.

10 Resolution within $h_3(\mathbb{O})$

A companion paper [**?**] argues that the universe's self-modeling algebra is the exceptional Jordan algebra $h_3(\mathbb{O})$, the unique 27-dimensional formally real Jordan algebra that is non-composable (it cannot participate in any composite with a nontrivial system). Within $h_3(\mathbb{O})$, the heuristic density ρ from Section 3 can be pinned down.

10.1 F_4-invariant measures

A normalized state $X \in h_3(\mathbb{O})$ (positive, $\mathrm{Tr}(X) = 1$) has three eigenvalues $\lambda_1, \lambda_2, \lambda_3 \geq 0$ summing to 1. The automorphism group $\mathrm{Aut}(h_3(\mathbb{O})) = F_4$ permutes eigenvectors while preserving eigenvalues. Any F_4-invariant function of a normalized state therefore depends only on the two free invariants:

$$\sigma_2 = \mathrm{Tr}(X^2) = \lambda_1^2 + \lambda_2^2 + \lambda_3^2, \qquad \sigma_3 = \det(X) = \lambda_1\,\lambda_2\,\lambda_3.$$

10.2 The uniqueness theorem

Theorem 8 (Uniqueness of ρ_J)**.** *Among non-negative F_4-invariant polynomials on the state space of $h_3(\mathbb{O})$ that vanish at rank-deficient states ($\sigma_3 = 0$) and at thermal equilibrium ($\sigma_2 = \frac{1}{3}$), the function*

$$\rho_J(X) \;=\; \det(X)\left(\mathrm{Tr}(X^2) - \tfrac{1}{3}\right) \tag{13}$$

is the unique candidate of minimal polynomial degree (degree 5 in the eigenvalues).

Proof. (i) F_4-invariance forces $\rho = f(\sigma_2, \sigma_3)$.

(ii) The zero at $\sigma_3 = 0$ forces divisibility: $\rho = \sigma_3 \cdot Q(\sigma_2, \sigma_3)$ for some polynomial Q.

(iii) The zero at thermal equilibrium ($\sigma_2 = \frac{1}{3}$, $\sigma_3 = \frac{1}{27}$) forces $Q(\frac{1}{3}, \frac{1}{27}) = 0$.

(iv) At degree 1 in (σ_2, σ_3), the general solution to (iii) is $Q = a\,(\sigma_2 - \frac{1}{3}) + b\,(\sigma_3 - \frac{1}{27})$. Since σ_2 has degree 2 and σ_3 has degree 3 in the eigenvalues, the a-term contributes degree $3 + 2 = 5$ and the b-term contributes degree $3 + 3 = 6$. At degree 5, the b-term vanishes: $b = 0$.

(v) Non-negativity. For $\mathrm{Tr}(X) = 1$ and $\lambda_i \geq 0$: $\sigma_3 \geq 0$ (product of non-negatives), and $\sigma_2 \geq \frac{1}{3}$ (by the QM-AM inequality applied to $\sum \lambda_i^2 \geq (\sum \lambda_i)^2/3 = \frac{1}{3}$). Therefore $\rho_J = \sigma_3\,(\sigma_2 - \frac{1}{3}) \geq 0$.

(vi) The only alternative at degree 1 with $a = 0$ is $\sigma_3 \cdot (\sigma_3 - \frac{1}{27})$. Since $\sigma_3 \in [0, \frac{1}{27}]$, this is non-positive everywhere. No valid measure.

Therefore ρ_J is the unique degree-5 candidate, up to positive scaling. □

Remark 9 (Connection to the information-theoretic formula). The heuristic $\rho = I(B;\, M)\,(1 - I/H)$ from Section 3 has the structure (correlation) $\times$ (distance from equilibrium). Within $h_3(\mathbb{O})$, the cubic norm $\det(X)$ measures three-way correlation between all Peirce sectors (the exceptional invariant), and $\mathrm{Tr}(X^2) - \frac{1}{3}$ measures purity above the maximally mixed baseline. These are the algebraic counterparts: $I(B;\, M) \leftrightarrow \det(X)$ and $1 - I/H \leftrightarrow \mathrm{Tr}(X^2) - \frac{1}{3}$. Same shape, same zeroes, different mathematical languages.

Remark 10 (The cubic prediction). ρ_J depends on $\det(X)$, a *cubic* invariant (the product of all three eigenvalues). This vanishes when *any* Peirce sector is empty. By contrast, integrated information Φ [**?**] is a quadratic measure (pairwise integration) that remains nonzero when one sector is disrupted as long as the other two remain connected. The prediction: disrupting one of three sectors (e.g., thalamocortical interface) should produce a *sharp* loss of self-modeling quality ($\det \to 0$), not a gradual reduction (Φ decreasing smoothly). This is the program's most distinguishing falsifiable prediction.

11 Discussion

11.1 What it does

The framework contributes:

- A control-theoretic motivation for self-modeling as a requirement for observerhood, grounded in Conant–Ashby and the internal model principle (Section 1.2).
- A quotienting principle that identifies copies by construction on $\mathcal{S}$ (definitional, conditional on P1).
- A computable density functional that discriminates self-modeling systems from non-self-modeling ones (demonstrated on toy model).
- A proof sketch for BB suppression (conditional on $\Delta_s > \Delta_b$) via exponential metastability, with numerical verification on a reduced system. A complete proof is given in [8].
- A concrete program for extending to "which structures are typical" once a reference measure is chosen.

A companion paper [7] proves that faithful self-modeling on a finite-dimensional spectral order-unit space forces the state space to carry C*-algebraic structure - complex quantum mechanics. The derivation chain proceeds: self-modeling $\to$ sequential product $\to$ Euclidean Jordan algebra [26] $\to$ local tomography (from faithful tracking) $\to$ complex type only (excluding real, quaternionic, spin factor, and exceptional Jordan algebras) $\to$ C*-algebra [27, 3, 16]. Complex numbers and the Born rule are outputs of this chain, not inputs. This sharpens ρ's role: the density selects self-modelers, and the algebraic consequences of self-modeling produce quantum mechanics by theorem.

11.2 What it does not

This doesn't explain consciousness. The density function measures self-modeling fidelity, which is correlated with consciousness in biological systems. Whether self-modeling *is* consciousness, *causes* it, or merely *co-occurs* with it is not addressed. The framework needs self-modeling to be *where consciousness lives* (the locative claim). Whether it's *what consciousness is* (the identity claim) is someone else's paper.

This doesn't derive $\Delta_s > \Delta_b$. The BB negligibility result is conditional on the basin depth gap. The gap is a physics input.

This doesn't construct a full measure on structure space. A universal prior ν over all mathematical structures is not constructed. The within-structure and within-universe results do not require it. Cross-structure typicality predictions would.

This doesn't resolve BBs in eternal inflation. Theorem A requires $\varepsilon \to 0$ (low noise) and finite T. Eternal inflation at fixed temperature is the opposite regime.

This doesn't extend to infinite-dimensional state spaces. The companion QM derivation [7] and the results here are stated for finite-dimensional systems. Extension to Type III von Neumann algebras (quantum field theory) is open.

The density is unique within $h_3(\mathbb{O})$ but not in general. Section 10 proves uniqueness at minimal polynomial degree within the exceptional Jordan algebra. On general self-modeling systems (not necessarily $h_3(\mathbb{O})$), other densities with the same qualitative shape could serve the same role.

This doesn't derive the Born rule from ρ. A separate investigation [6] tested whether ρ itself derives Born-rule measurement probabilities. It does not: $\rho_Q \leq 0$ throughout Lindblad evolution. This is now understood as confirmation, not failure. The Born rule follows from Gleason's theorem [14] applied to the C*-algebra that self-modeling forces [7]. The density ρ selects self-modelers; Gleason handles measurement probabilities.

11.3 Open problems

Coarse-graining invariance. The density depends on the representational level. State padding (appending independent noise dimensions to B) increases $H(B)$ while leaving $I(B;\, M)$ fixed, shifting ρ toward its maximum. The Markov-level axiom (P2) blocks this within the framework by declaring the coarse-graining fixed. But a theory on structure space should be invariant under information-neutral refinements and padding. It isn't yet. Replacing $H(B)$ with excess entropy or predictive information in B alone would close this gap. Until it's closed, the density is representation-sensitive in a way that limits the theory's claims about structure-space invariance.

Why $I/H \approx 0.5$ for biological observers. Section 1.2 argues that evolution concentrates self-modeling systems near the peak of ρ. The qualitative argument is observational; the quantitative prediction – that biological self-models cluster near $I/H = 0.5$ specifically – requires verification via evolutionary game theory simulation.

Strengthening the self-model criterion. The factorization condition detects tracking structure. It doesn't distinguish genuine self-modeling from passive readout, correlated aux-

iliary registers, or memory traces that serve no regulatory function. This is the framework's most important technical gap after coarse-graining invariance. The minimum strengthening: (i) the model must track body variables, (ii) the model must causally influence body evolution (bidirectional coupling, not just observation), and (iii) that influence must improve persistence or viability relative to a counterfactual without it. This would bring the formalism closer to the Conant–Ashby motivation. Exact enumeration over product decompositions is also intractable for large systems; the strengthened criterion would narrow the search.

Reference measure on $\mathcal{S}$. A universal prior ν over structure space would extend the framework from within-universe claims to cross-structure typicality. Candidate approaches: algorithmic probability, maximum-entropy over kernels, or a physics-derived ensemble. Note that structures with $W = 0$ contribute nothing to the integral regardless of ν, restricting the effective support to the subspace of $\mathcal{S}$ with nontrivial self-modeling.

Cross-structure BBs. The within-structure BB result (Theorem A) suppresses BBs that are excursions within a single chain. The quotient collapses exact duplicates across chains. What remains are *distinct* composite processes – different points in $\mathcal{S}$ – with nonzero ρ but unstructured body dynamics. These systems may carry nontrivial experiential weight, but their lack of temporal body structure means $\rho_{\text{pred}} \approx 0$: the self-model cannot predict the body's future. Whether the $\rho_{\text{pred}} \approx 0$ region of $\mathcal{S}$ dominates the $\rho_{\text{pred}} \gg 0$ region under natural choices of ν is an open question that would close the cross-structure BB problem.

Infinite-dimensional extension. The companion QM derivation [7] and the results here are stated for finite-dimensional systems. Extending to Type III von Neumann algebras (quantum field theory) requires new techniques: the sequential product axioms and the van de Wetering classification [26] both assume finite dimension. Whether the self-modeling forcing argument survives in the infinite-dimensional setting is the principal open mathematical question for the program.

Gravity. The companion result [7] derives the internal algebra (QM) from self-modeling. It does not derive the external geometry (GR). If gravity emerges from the information-geometric structure of self-modeling - the locality bottleneck through which a finite subsystem accesses a larger quantum system - this would connect to Jacobson's thermodynamic derivation of Einstein's equations [18] and Verlinde's entropic gravity program. This remains a conjecture.

11.4 Summary of status

This paper constructs an experiential measure for self-modeling systems: a density functional ρ, a structure space $\mathcal{S}$ with groupoid quotient, and a conditional proof for Boltzmann brain suppression (fully proved in [8]). The core claims are within-structure and do not require a universal prior.

A companion result [7] proves that faithful self-modeling forces the state space to carry C*-algebraic structure (complex quantum mechanics), with the Born rule following from

Gleason's theorem. This upgrades ρ from a candidate selection criterion to the selection half of a two-part story: ρ selects self-modelers, and self-modeling forces quantum mechanics. The principal open problems are the reference measure ν on structure space (extending the framework to cross-structure typicality predictions), the infinite-dimensional extension (QFT), and the gravity conjecture.

References

[1] N. Ackerman, C. Freer, and R. Patel. Invariant measures concentrated on countable structures. *Forum of Mathematics, Sigma*, 2016.

[2] J. C. Baez and J. Dolan. From finite sets to Feynman diagrams. In *Mathematics Unlimited – 2001 and Beyond*, Springer, 2001.

[3] H. Barnum and A. Wilce. Post-classical probability theory. In G. Chiribella and R. W. Spekkens, editors, *Quantum Theory: Informational Foundations and Foils*, Springer, 2016.

[4] A. Bovier, M. Eckhoff, V. Gayrard, and M. Klein. Metastability in reversible diffusion processes I: Sharp asymptotics for capacities and exit times. *J. Eur. Math. Soc.*, 2004.

[5] S. M. Carroll. Why Boltzmann brains are bad. In *Current Controversies in Philosophy of Science*, Routledge, 2017.

[6] B. Ehrlich. Born-rule distributions and self-modeling information: a negative result. *ehrlich.dev/papers/born-fisher-2026.pdf*, 2026.

[7] B. Ehrlich. Quantum mechanics from self-modeling: one premise. In preparation, 2026.

[8] B. Ehrlich. Self-contained proof of Theorem A: Boltzmann brain negligibility via metastability. *ehrlich.dev/papers/theorem-a-proof.pdf*, 2026.

[9] R. C. Conant and W. R. Ashby. Every good regulator of a system must be a model of that system. *Int. J. Systems Sci.*, 1(2):89–97, 1970.

[10] A. L. Foote and J. D. Crystal. Metacognition in the rat. *Current Biology*, 2007.

[11] D. S. Freed, M. J. Hopkins, J. Lurie, and C. Teleman. Topological quantum field theories from compact Lie groups. In *A Celebration of the Mathematical Legacy of Raoul Bott*, CRM Proceedings. AMS, 2010.

[12] B. A. Francis and W. M. Wonham. The internal model principle of control theory. *Automatica*, 12(5):457–465, 1976.

[13] M. I. Freidlin and A. D. Wentzell. *Random Perturbations of Dynamical Systems*. 3rd ed., Springer, 2012.

[14] A. M. Gleason. Measures on the closed subspaces of a Hilbert space. *J. Math. Mech.*, 6(6):885–893, 1957.

[15] M. Gell-Mann and S. Lloyd. Information measures, effective complexity, and total information. *Complexity*, 1996.

[16] H. Hanche-Olsen. On the structure and tensor products of JC-algebras. *Canad. J. Math.*, 35(6):1059–1074, 1984.

[17] R. R. Hampton. Rhesus monkeys know when they remember. *Proc. Natl. Acad. Sci.*, 2001.

[18] T. Jacobson. Thermodynamics of spacetime: the Einstein equation of state. *Phys. Rev. Lett.*, 75(7):1260–1263, 1995.

[19] R. Lopez-Ruiz, H. L. Mancini, and X. Calbet. A statistical measure of complexity. *Physics Letters A*, 1995.

[20] M. P. Müller. Law without law: from observer states to physics via algorithmic information theory. *Quantum*, 2020.

[21] M. P. Müller. Algorithmic idealism: What should you believe to experience next? *Foundations of Physics*, 56:11, 2026.

[22] A. A. Pinto and D. A. Rand. Classifying C^{1+} structures on dynamical fractals: 2. Embedded trees. *Ergodic Theory Dynam. Systems*, 2006.

[23] M. Tegmark. Is "the theory of everything" merely the ultimate ensemble theory? *Annals of Physics*, 1998.

[24] M. Tegmark. The mathematical universe. *Foundations of Physics*, 2008.

[25] G. Tononi et al. Integrated information theory: from consciousness to its physical substrate. *Nature Reviews Neuroscience*, 2016.

[26] J. van de Wetering. An effect-theoretic reconstruction of quantum theory. *Compositionality*, 1:1, 2019.

[27] J. van de Wetering. Sequential product spaces are Jordan algebras. *J. Math. Phys.*, 60(6):062201, 2019.

[28] A. Weinstein. The volume of a differentiable stack. *Lett. Math. Phys.*, 2009.

[29] D. H. Wolpert and J. Kipper. Memory systems, the epistemic arrow of time, and the second law. *Entropy*, 26(2):170, 2024.

A Toy model code

Full source code available at: `toy-model/composite_self_model.py` (seven composite-state Markov chains), `bipartition_test.py` (partition search), `predictive_mi.py` (predictive MI computation), `trajectory_demo.py` (trajectory-level metastability verification), `three_state_chain.py` (BB negligibility verification).

B Lemma dependency graph

```
L1 (Basin Partition via FW cycle hierarchy)
 |-- L2 (Exponential Residence Time - BEGK)
 |    |-- L4 (Experiential Measure Lower Bound, B_stable)
 |    +-- L5 (BB Occupation Time Upper Bound)
 |-- L3 (QSD Convergence within B_stable)
 |    +-- L4
 +-- L6 (DV Concentration for Weighted Empirical Measure)
      +-- L7 (Ratio Bound Assembly)
              ^
       L4, L5 -+
```

C Failure modes and resolution status

Table 4 summarizes the adversarial tests applied to the framework and their resolution status.

D Mathematical details

D.1 Continuous-time formulation

CTMC analogue (bipartite jump structure). In continuous time, "update order" is not primitive - the generator Q on $\Omega = B \times M$ encodes instantaneous rates, not sequential steps. The corresponding structural condition is that Q admits a **bipartite jump structure**: every off-diagonal entry $Q\big((b,m),(b',m')\big)$ with $(b',m') \neq (b,m)$ satisfies either $b' \neq b$ and $m' = m$ (a body jump) or $b' = b$ and $m' \neq m$ (a model jump). Simultaneous jumps - transitions where both b and m change - have rate zero. Body jump rates $Q_B\big((b,m) \to (b',m)\big)$ may depend on the full state (b,m). Model jump rates $Q_M\big((b,m) \to (b,m')\big)$ may also depend on (b,m): the model's rate can depend on the current body state, which is the infinitesimal version of "model observes body."

Relating the two formulations. The discrete-time kernel $P = P_B \cdot P_M$ is a *composition* of two substep kernels in which both components may change per time step. The CTMC bipartite structure allows only one component to change per jump event. These are not the same model, but they are related: uniformizing Q (setting $\lambda = \max_x |Q(x,x)|$, forming $P_{\text{unif}} = I + Q/\lambda$) produces a discrete chain where only one component changes per step. The two-substep composition $P = P_B \cdot P_M$ is recovered by grouping pairs of uniformized steps (one B-jump followed by one M-jump) into a single macro-step. In the continuous-time limit (step size $\to 0$), both formulations converge to the same generator Q. The toy model uses the composition formulation with fixed physical time step $\mathrm{d}t = 1$; the formal framework uses the generator Q directly.

D.2 Structure space topology

Even for fixed $|\Omega| = |B|\cdot|M|$, the set of factorized transition kernels is a continuous parameter space: each kernel is a point in a product of simplices (one $|\Omega|$-simplex per row), forming a compact convex subset of $\mathbb{R}^{|\Omega|^2}$. The symmetry group $G = \mathrm{Sym}(B) \times \mathrm{Sym}(M)$ of product-preserving permutations is finite ($|G| = |B|!\,|M|!$) and acts on this space by conjugation. The quotient $\mathcal{S} = \{\text{factorized kernels}\}/G$ is therefore an orbifold on the free-action locus; globally a stratified space with boundary. Generic orbits (kernels with trivial stabilizer) form the top stratum; kernels with nontrivial automorphism groups sit on lower-dimensional strata.

We avoid calling this a "moduli space" without qualification. For restricted classes of smooth dynamical systems (e.g., hyperbolic basic sets), moduli-type parameterizations exist [22]. For the general case, the quotient may lack the nice topological properties that "moduli space" connotes. We use "structure space" to flag this distinction.

D.3 Groupoid cardinality derivation

Proposition 11 ($1/|\mathrm{Aut}|$ from G-invariant measures)**.** *Let $G = \mathrm{Sym}(B) \times \mathrm{Sym}(M)$ act on the space $\mathcal{K}$ of factorized kernels by conjugation. For any G-invariant base measure on $\mathcal{K}$, the induced weighting on the quotient $\mathcal{S} = \mathcal{K}/G$ is proportional to $1/|\mathrm{Aut}(P)|$. This holds in two regimes:*

Proof sketch. Counting/discrete measures. For a counting measure on a finite or countable subset of $\mathcal{K}$, the orbit-stabilizer theorem gives $|\mathrm{orbit}(P)| = |G|/|\mathrm{Stab}(P)| = |G|/|\mathrm{Aut}(P)|$. A G-invariant counting measure assigns equal weight to each element, so the total weight of orbit $[P]$ is proportional to $|G|/|\mathrm{Aut}(P)|$. Since $|G|$ is constant, relative weight is $1/|\mathrm{Aut}(P)|$.

Continuous measures. For the product Lebesgue measure λ on the simplex factors, individual orbits are finite sets and have λ-measure zero. In the ordinary pushforward to the quotient topological space, stabilizer size is invisible. The $1/|\mathrm{Aut}|$ factor only appears under **groupoid/stack integration**: the integral of a function f over $\mathcal{S}$ against a G-invariant base measure λ on $\mathcal{K}$ is defined as $(1/|G|)\int_{\mathcal{K}} f([P])\,\mathrm{d}\lambda(P)$. This is equivalent to integrating f on the quotient with the orbifold volume form, which carries a local $1/|\mathrm{Stab}(P)| = 1/|\mathrm{Aut}(P)|$ density factor [28]. Under this convention, $1/|\mathrm{Aut}|$ weighting is built into the definition of integration on the quotient, not derived from a pushforward. □

For absolutely continuous base measures on $\mathcal{K}$, the singular strata (kernels with nontrivial Aut) have positive codimension and therefore Lebesgue-measure zero. In ordinary pushforward integration, stabilizer size is invisible. The stack/groupoid convention makes it visible. This distinction matters primarily when the base measure has atoms (the counting/discrete case) or when performing local orbifold integration near singularities. The practical relevance of the $1/|\mathrm{Aut}|$ factor therefore depends on the type of reference measure ν: for a discrete Solomonoff-type prior over computable descriptions, the factor is substantive; for a smooth maximum-entropy prior over kernels, it affects only measure-zero strata.

What the quotient does and does not do. The groupoid quotient collapses exact duplicates: N copies of the same system in different spatial locations are one point on $\mathcal{S}$, not

N points. This eliminates the counting infinity that drives the BB problem in spacetime-based measures.

What it does *not* do is suppress distinct, asymmetric thermal fluctuations. A generic thermal microstate has trivial automorphism group, same as a brain. The $1/|\mathrm{Aut}|$ weighting doesn't help here. For distinct BB types, suppression comes from the density and metastability mechanisms (Sections 3 and 7).

D.4 Coarse-graining vulnerability

State padding. The normalization by $H(B)$ creates a specific attack: appending independent noise dimensions to B increases $H(B)$ while leaving $I(B;\, M)$ fixed, pushing ρ toward I (its maximum). The factorization condition and "coarsest factorization" selection rule do not block this, because enlarging B with independent noise preserves the factorization. The Markov-level axiom (P2) declares the coarse-graining fixed, which prevents the attack *within* the framework, but for a theory on structure space, invariance under representational padding is desirable. A density defined via a body entropy that excludes independent noise – e.g., excess entropy or predictive information in B alone – would close this gap. We flag state padding, along with state splitting (refining Ω while preserving macro-dynamics) and rate rescaling (Section D.11), as coarse-graining attacks that a mature version of the framework must address.

D.5 Admissible density family

Remark 12 (Admissible density family). The key results (Theorem A, copy collapse, toy-model discrimination) depend on qualitative properties of ρ, not on the specific parabolic form. An **admissible density** is any functional $\rho\colon \Delta(\Omega) \to [0, \infty)$ satisfying: (i) $\rho(p) = 0$ when $I_p(B;\, M) = 0$; (ii) $\rho(p) = 0$ when $I_p(B;\, M) = H_p(B)$; (iii) ρ is bounded; (iv) ρ is continuous in p; (v) ρ is invariant under product-preserving isomorphism. The $I(1 - I/H)$ form is the canonical example – the simplest admissible density – but any member of this family would produce the same qualitative BB suppression under Theorem A's hypotheses.

D.6 Approximate factorization and stability

Approximate factorization. Exact conditional independence equalities are measure-zero in the space of kernels. For realistic systems, the factorization will hold approximately. The formal relaxation: for a candidate decomposition (B, M), define the factorization gap

$$\delta(P, B, M) \;=\; \max_{b,\, b',\, m} \mathrm{KL}\big(P_M(\cdot \mid b',\, b,\, m) \;\|\; P_M(\cdot \mid b',\, m)\big) \tag{14}$$

where $P_M(\cdot \mid b', b, m)$ is the true conditional on M's next state given full history, and $P_M(\cdot \mid b', m)$ is obtained by marginalizing over b with weights from the stationary distribution $\pi(b \mid m)$. When $\delta = 0$, the conditional independence (Eq. 3) holds exactly. If $\delta <$ threshold, the decomposition is "approximately factorized."

Selection principle. When multiple approximate decompositions exist, maximizing ρ over all of them risks overfitting. The selection rule: among all decompositions with $\delta(P, B, M) <$ threshold, select the **coarsest** factorization (largest $|B|$, smallest $|M|$). This corresponds to a minimum description length criterion: prefer the decomposition that explains the self-modeling structure with the smallest model component, analogous to penalizing model complexity in statistical learning. Making this explicit as a regularized score $\text{score}(B, M) = \rho - \lambda \log |M| - \kappa\, \delta$ under monotonicity assumptions on ρ vs. $|M|$ is deferred.

Conjecture 13 (Lipschitz stability)**.** *If P and P' are two kernels on Ω with $\|P - P'\|_\infty < \eta$, and both admit the same (B, M) decomposition with factorization gaps below threshold, then $|\rho(P) - \rho(P')| \leq L\, \eta$, where L depends on $|\Omega|$ and the spectral gap of P.*

The expected argument: (i) the stationary distribution π satisfies $\|\pi - \pi'\|_1 \leq (1/\gamma)\|P - P'\|_\infty$ where γ is the spectral gap; (ii) mutual information is Lipschitz in total variation on distributions bounded away from zero; (iii) composing (i) and (ii) gives the claim. Proof with explicit constants is deferred.

D.7 Trajectory functional details

Axiom (Markov level). The composite Markov process (Ω, B, M, Q) specifies a complete dynamical description at a fixed level of coarse-graining. The distribution p_t over (B, M) is the state of the system at this level - not epistemic uncertainty about a finer microstate, but the operative description at which self-modeling structure is defined. The experiential density $\rho(p_t)$ is a property of this level.

This is analogous to how statistical mechanics treats the Boltzmann distribution: not as ignorance about which microstate a gas occupies, but as the macroscopic description at the thermodynamic level. The analogy is imperfect - Boltzmann distributions are typically justified by typicality or ergodicity, which is additional structure we do not invoke. Our commitment is simpler: the Markov level at which the factorization (B, M) is defined is the level at which ρ is evaluated.

Remark 14. An alternative path-level quantity: define $\hat{\rho}(x) = \rho(\delta_x)$ for individual states x and take $\mu_{\text{path}} = \mathbb{E}[\int \hat{\rho}(X_t)\, \mathrm{d}t]$ under the path measure. For finite chains, $\mathbb{E}[\int \hat{\rho}(X_t)\, \mathrm{d}t] \neq \int \rho(p_t)\, \mathrm{d}t$ in general ($\rho = I(1 - I/H)$ has mixed convexity properties and the product form inherits no clean convexity). The distributional version is what we mean by μ. The path-level version is a different functional.

Time-rate sensitivity. The continuous-time definition eliminates the time-refinement vulnerability, but Q and cQ (uniform speedup by factor $c > 0$) produce different trajectory integrals: $\mu(cQ, [0, T]) = c\, \mu(Q, [0, T])$ at stationarity. A system running at twice the rate accumulates twice the experiential measure. This is physically meaningful: a system with faster transition rates processes more information per unit external time.

For the BB problem specifically, the comparison in Section 7 is between excursions *within a single rate matrix Q*. The speedup concern does not apply within-structure.

For cross-structure comparison, one can define the rate-normalized functional $\bar{\mu}([0, T]) = \int_0^T \rho(p_t)/\lambda(Q)\, \mathrm{d}t$, where $\lambda(Q) = \max_x |Q(x, x)|$ is the uniformization rate. This makes $\bar{\mu}$

dimensionless and invariant under $Q \to cQ$. We do not adopt this normalization here because within-structure comparisons do not require it.

D.8 Toy model construction details

Metastable body. Two basins $\{0, 1\}$ and $\{2, 3\}$ with rare cross-basin transitions:

$$P_B = \begin{pmatrix} 0.70 & 0.25 & 0.04 & 0.01 \\ 0.25 & 0.70 & 0.01 & 0.04 \\ 0.04 & 0.01 & 0.70 & 0.25 \\ 0.01 & 0.04 & 0.25 & 0.70 \end{pmatrix}. \tag{15}$$

Within-basin transitions (e.g., $0 \leftrightarrow 1$) have probability 0.25; cross-basin transitions (e.g., $0 \to 2$) have probability 0.04 or 0.01.

Cyclic body. Near-deterministic cycle $0 \to 1 \to 2 \to 3 \to 0$:

$$P_B = \begin{pmatrix} 0.01 & 0.97 & 0.01 & 0.01 \\ 0.01 & 0.01 & 0.97 & 0.01 \\ 0.01 & 0.01 & 0.01 & 0.97 \\ 0.97 & 0.01 & 0.01 & 0.01 \end{pmatrix}. \tag{16}$$

Random body. Near-uniform: $P_B(b' \mid b) \approx 1/4$ for all b, b', with random noise ± 0.03 to break exact uniformity.

Model update rule. For tracking models, the update given new body state b' and old model state m:

$$P_M(m' \mid b',\, m) \;=\; \alpha\, \delta_{m',b'} \;+\; \beta\, \delta_{m',m} \;+\; \frac{\gamma}{|M|}$$

where α is tracking accuracy, $\beta = \max(1 - \alpha - \gamma,\, 0)$ is persistence, and $\gamma = 0.02$ is uniform noise. For independent models, $P\big((b', m') \mid (b, m)\big) = P_B(b' \mid b) \cdot P_M(m' \mid m)$.

Worked example. Observer (system 1), $\alpha = 0.5$, $\gamma = 0.02$, $\beta = 0.48$. At state $(b, m) = (0, 2)$, if the body transitions to $b' = 1$:

$$P_M(m' \mid b' = 1,\, m = 2) \;=\; \begin{cases} 0.505 & m' = 1 \text{ (track new body)} \\ 0.485 & m' = 2 \text{ (persist)} \\ 0.005 & m' = 0,\, 3 \text{ (noise)} \end{cases}$$

Full joint transition: $P_B(1 \mid 0) \cdot P_M(1 \mid 1, 2) = 0.25 \times 0.505 = 0.126$.

D.9 Three-state chain details

The three-state chain (states s, u, b arranged as $s \leftrightarrow u \leftrightarrow b$) uses Freidlin–Wentzell transition rates:

$$P(s \to u) = e^{-\Delta_s/\varepsilon}, \qquad P(u \to s) = p,$$
$$P(b \to u) = e^{-\Delta_b/\varepsilon}, \qquad P(u \to b) = 1 - p,$$

and self-loops $P(s \to s) = 1 - e^{-\Delta_s/\varepsilon}$, $P(b \to b) = 1 - e^{-\Delta_b/\varepsilon}$.

The stationary distribution from detailed balance:

$$\pi_s = \frac{p\, e^{\Delta_s/\varepsilon}}{Z}, \qquad \pi_u = \frac{1}{Z}, \qquad \pi_b = \frac{(1-p)\, e^{\Delta_b/\varepsilon}}{Z},$$

where $Z = p\, e^{\Delta_s/\varepsilon} + 1 + (1-p)\, e^{\Delta_b/\varepsilon}$.

D.10 Formal theorem statement

Theorem 15 (Theorem A: metastability-based BB negligibility)**.** ***Setup.*** *Let (Ω, Q_ε) be a family of finite irreducible continuous-time Markov chains indexed by $\varepsilon > 0$ (noise intensity), with rate matrices corresponding to reversible Metropolis-type dynamics:*

$$Q_\varepsilon(x, y) \;\propto\; \exp\!\big(-[E(y) - E(x)]^+/\varepsilon\big)$$

for an energy function E on Ω. Let $\mathcal{B}_{\text{stable}}$ be a metastable set with communication height Δ_s and $\mathcal{B}_{\text{BB}}$ be a metastable set with communication height Δ_b, with $\Delta_s > \Delta_b$ (communication height = Freidlin–Wentzell quasipotential depth). Let $\rho(p_t)$ be any admissible density (Appendix D.5) satisfying $\rho(p_t) \geq c > 0$ at the QSD of $\mathcal{B}_{\text{stable}}$.

Statement. *For trajectories over $T_\varepsilon = \exp\big((\Delta_s - \alpha)/\varepsilon\big)$ starting from $\mathcal{B}_{\text{stable}}$ (where $0 < \alpha < \Delta_s - \Delta_b$):*

$$\frac{\mu_{\text{BB}}}{\mu_{\text{stable}}} \;\leq\; C \cdot \exp\!\Big(-\frac{\Delta_s - \Delta_b - \alpha}{\varepsilon}\Big) \;\to\; 0 \quad \text{as } \varepsilon \to 0.$$

Scope. This is a metastability result for reversible finite-state Markov chains. Real observers are non-equilibrium dissipative structures; extending to broken detailed balance requires non-reversible BEGK machinery. This result does not by itself resolve BB problems in cosmological models with infinite future time at fixed temperature.

D.11 Time-discretization invariance

The continuous-time integral μ is invariant under time refinement: the rate matrix Q determines a unique flow, and μ depends only on Q and T, not on any discretization. A discrete-time formulation $\mu_T = \sum \rho_t$ would be vulnerable to inflating the sum by inserting identity steps. The toy model uses discrete-time chains with a fixed physical time step $\mathrm{d}t = 1$; the discrete sum is a Riemann approximation to the continuous-time integral.

D.12 Proof details for Theorem A

Lemma decomposition. The proof decomposes into seven lemmas (dependency graph in Appendix B):

1. **Basin partition** (FW cycle hierarchy): Decompose state space into deep and shallow basins.
2. **Residence time lower bound** (BEGK potential theory): The system stays in $\mathcal{B}_{\text{stable}}$ for expected time $\sim \exp(\Delta_s/\varepsilon)$.
3. **QSD convergence** (Doob h-transform, spectral theory): Within $\mathcal{B}_{\text{stable}}$, the trajectory distribution converges exponentially fast to the quasi-stationary distribution.
4. **Stable measure lower bound**: Combining (2) and (3), the experiential measure accumulated in $\mathcal{B}_{\text{stable}}$ is $\geq c\,c'\,\exp\big((\Delta_s - \alpha)/\varepsilon\big)$.
5. **BB occupation time upper bound** (renewal theory): Before exiting $\mathcal{B}_{\text{stable}}$, the expected total time in BB states scales as $\exp(\Delta_b/\varepsilon)$.
6. **DV concentration** (Donsker–Varadhan large deviations): The weighted empirical measure concentrates around its expected value.
7. **Ratio assembly**: Combining (4) and (5) with boundary vanishing gives the exponential bound.

All tools used - Freidlin–Wentzell theory [13], BEGK potential theory [4], Donsker–Varadhan large deviations, Champagnat–Villemonais QSD theory - are mature and well-established. The individual lemmas follow from known results; what remains is assembling them into a complete proof with explicit constants and verifying that the composition of error bounds closes. This is a program, not a completed proof.

Complementary theorem ($T \to \infty$). For the $T \to \infty$ regime, negligibility holds but requires the *joint* limit $\varepsilon \to 0$ and $T \to \infty$. In the renewal theory extension (trajectory alternates between basins), the asymptotic ratio is

$$\frac{\delta \cdot \exp(\Delta_b/\varepsilon)}{c \cdot \exp(\Delta_s/\varepsilon)} \;=\; \frac{\delta}{c}\,\exp\Big(-\frac{\Delta_s - \Delta_b}{\varepsilon}\Big) \;\to\; 0.$$

Failure mode	Severity	Status	Resolution
Aaronson XOR grid	Medium	Resolved	Under dynamically realized partition, crystal/XOR has $I/H \to 1$, $\rho \to 0$. Under unrestricted max-over-partitions, crystal scores as high as observer (Section 5.8).
Partition dependence	High	Resolved	Factorization condition identifies partition from kernel (Section 3.7).
Flat landscape ($\Delta_s \approx \Delta_b$)	High	Open	Conditional theorem. $\Delta_s > \Delta_b$ is a stated assumption.
Perfect duplicate BB	High	Addressed	Quotient collapses isomorphic systems, conditional on P1 (Section 8).
BB + self-model	Medium	Reframed	Non-zero ρ is correct. Suppression via trajectory integral (Section 5.5).
Reference measure gap	High	Open	Full measure requires prior ν (Section 2.5).
ρ type error	High	Resolved	ρ is functional of time-marginal distribution, not state-local (Section 3.1).
CTMC factorization	Medium	Resolved	Bipartite jump structure at generator level (Section 2.1).
Representational vs. predictive MI	Medium	Co-equal	Both ρ and ρ_{pred} provide independent suppression channels (Section 3.6).
Time-discretization	Critical	Resolved	μ defined as continuous-time integral. Time-refinement vulnerability eliminated (Section D.11).
Approximate factorization	High	Formalized	KL-divergence relaxation with coarsest-first selection (Section 3.7).
Proof completeness	High	Scoped	Theorem A is a proof sketch, not a completed proof (Section 1.5).
Claim strength	High	Re-scoped	Universal language replaced with conditional/scoped language (Section 1.5).

Table 4: Adversarial failure modes applied to the framework.