

Substrate Theory: A Formally Verified Complexity-Threshold Framework for Modeling Informational Regimes in Physics

[Anonymized for Review]

Abstract

We present a formally verified framework—“Substrate Theory”—for studying how informational complexity thresholds may delimit distinct computational regimes relevant to physical modeling. The framework separates (i) an *ideal layer* based on noncomputable Kolmogorov complexity, (ii) an *operational layer* based on a computable proxy (here, Lempel–Ziv complexity), and (iii) a *bridge layer* establishing provable relationships between the two. The central construct is a *grounding threshold parameter* C_{ground} that is treated explicitly as a free, empirically constrainable quantity rather than a derived constant. Below C_{ground} , the operational rules preserve informational coherence (reversible, history-preserving behavior); above C_{ground} , they permit information-reducing updates. The entire system is mechanized in the Lean 4 theorem prover, providing machine-checked internal consistency of definitions and theorems. We do *not* claim a unification of existing physical theories nor derivations of fundamental constants from first principles. Instead, we offer a logically precise scaffold on which such hypotheses can be formulated, compared, and tested against experiment.

Keywords: algorithmic information; Kolmogorov complexity; computable proxies; Lean 4; mechanized reasoning; coherence vs. reduction; empirical thresholds.

Supplementary material. A complete canonical specification (Lean 4 source and formal definitions/proofs) is provided as anonymised supplementary material accompanying this submission.

1 Introduction: Why Complexity Thresholds?

Foundational questions in physics—from the quantum-classical transition to the emergence of spacetime—increasingly invite information-theoretic perspectives.¹ However, a persistent challenge lies in bridging the gap between abstract algorithmic notions and concrete operational measurements. How can we rigorously test hypotheses about complexity-driven regime transitions when our fundamental complexity measures are noncomputable?

This work addresses this challenge by developing a *formally verified* framework for studying complexity-threshold dynamics. Our key insight is to separate concerns across three distinct layers:

- An *ideal layer* using noncomputable Kolmogorov complexity as a conceptual baseline
- An *operational layer* using computable proxies (like Lempel–Ziv) for practical measurements

¹For background, see, e.g., Kolmogorov [1] on algorithmic complexity; Lempel–Ziv [2] on practical coding; and information-theoretic approaches to decoherence [3] and statistical mechanics.

- A *bridge layer* with machine-checked theorems connecting the two

The central innovation is treating the regime boundary C_{ground} as an explicit *free parameter* to be determined empirically, rather than deriving it from first principles. This honest approach allows the framework to serve as a testable interface between algorithmic information theory and physical phenomena.

Example 1.1 (Toy Model: Coherent vs. Random Dynamics). *Consider a simple cellular automaton where patterns below a certain complexity threshold exhibit coherent, predictable evolution (akin to quantum unitary dynamics), while highly complex patterns undergo information-reducing collapses (akin to measurement). Our framework provides the formal structure to define such thresholds precisely and verify that the operational rules maintain desired informational properties.*

Our contribution is methodological: we provide a machine-checked scaffold for formulating and testing complexity-threshold hypotheses. The entire development is mechanized in Lean 4, ensuring logical consistency. We make *no* claims about unifying existing physical theories; rather, we offer a disciplined way to explore such possibilities.

2 Overview of the Three-Layer Architecture

2.1 Ideal Layer: The Noncomputable North Star

At the ideal layer, we use Kolmogorov complexity $K(\cdot)$ and its conditional and joint variants as abstract baselines. While noncomputable, K provides the “conceptual north star” for what complexity-theoretic statements *should* mean. For instance, the conditional complexity $K(x|y)$ measures the information in x not already present in y , formalizing notions of emergence and grounding.

2.2 Operational Layer: Computable Proxies

Since K is noncomputable, we replace it with computable proxies K_{LZ} in the operational layer. We primarily use Lempel–Ziv complexity, but the framework is proxy-agnostic. The operational layer supports executable rules and measurable predictions, making it suitable for experimental testing.

2.3 Bridge Layer: Provable Connections

The bridge layer contains theorems that bound K_{LZ} in terms of K (up to additive/multiplicative constants) and propagate structural properties from ideal to operational layers. These results justify using K_{LZ} for regime tests while maintaining traceability to K -level concepts.

3 Core Framework Elements

3.1 Basic Definitions

Definition 3.1 (States, Entities, Substrate). Let `State` be finite bitstrings (lists of Booleans). Let `Entity` be an abstract type with encodings into `State`. The *substrate* is a distinguished minimal-complexity entity serving as an informational baseline.

Definition 3.2 (Complexity Measures). For $x \in \text{Entity}$, $K(x)$ denotes prefix Kolmogorov complexity (ideal, noncomputable), while $K_{\text{LZ}}(x)$ denotes Lempel–Ziv complexity (operational, computable). Conditional and joint variants are defined analogously.

3.2 The Central Threshold Parameter

Axiom 3.1 (Operational Regimes via C_{ground}). There exists a parameter $C_{\text{ground}} \in \mathbb{N}$ such that:

- States with $K_{\text{LZ}}(\cdot) \leq C_{\text{ground}}$ are *coherence-preserving* (reversible, history-keeping)
- States with $K_{\text{LZ}}(\cdot) > C_{\text{ground}}$ admit *information-reducing* updates

The value of C_{ground} is *not derived* but must be empirically determined for each application domain.

This threshold parameter C_{ground} serves as the crucial empirical interface between the formal framework and physical reality.

3.3 Dynamical Rules

Definition 3.3 (Rule Families). We define two update families on state histories:

- R_{coh} : History-preserving (reversible) update for low-complexity regimes ($K_{\text{LZ}} \leq C_{\text{ground}}$)
- R_{red} : History-forgetting (information-reducing) update for high-complexity regimes ($K_{\text{LZ}} > C_{\text{ground}}$)

4 Key Formal Results

We summarize representative theorems; complete machine-checked proofs are available in the supplementary Lean code.

Theorem 4.1 (Proxy Calibration). *There exist constants $a, b \geq 0$ such that for encoded states x, y :*

$$K_{\text{LZ}}(xy) \leq K_{\text{LZ}}(x) + K_{\text{LZ}}(y) + a,$$

and under standard mixing conditions:

$$|K_{\text{LZ}}(x) - K(x)| \leq b + o(|x|).$$

This ensures the computable proxy K_{LZ} approximates the ideal K within bounded error.

Theorem 4.2 (Regime Stability). • *Below threshold: Coherence rules preserve information measures*

- *Above threshold: Reduction rules guarantee bounded information loss*

Formally, for histories h in the low-complexity regime, application of R_{coh} preserves coherence functionals, while R_{red} ensures $K_{\text{LZ}}(R_{\text{red}}(h)) \leq K_{\text{LZ}}(h) + c_{\text{over}} - c_{\text{drop}}$ with $c_{\text{drop}} > 0$.

These theorems certify that the operational rules respect the intended informational design, providing formal guarantees for domain applications.

5 From Formal Framework to Physical Modeling

To prevent conflation of logic with interpretation, we strictly separate *logical theorems* (proved within the framework) from *physical postulates* (interpretive assumptions).

5.1 Physical Interface Postulates

- P1. **Encoding Discipline:** Physical configurations admit reproducible encodings into State
- P2. **Threshold Hypothesis:** Observable regime boundaries correlate with K_{LZ} measurements at some C_{ground}
- P3. **Dynamics Mapping:** Physical dynamics correspond to R_{coh} below C_{ground} and R_{red} above C_{ground}

These postulates form the testable interface between the formal framework and physical reality. They are *not* consequences of the mathematics but must be validated empirically.

6 Testing the Framework: A Detailed Protocol

The framework enables concrete, falsifiable testing through a three-stage protocol:

6.1 Stage 1: Proxy Calibration

1. **Encode:** Map physical observables to bitstrings via reproducible encoding pipelines
2. **Measure:** Compute K_{LZ} statistics on appropriately preprocessed data
3. **Correlate:** Identify whether qualitative behavior transitions correlate with stable K_{LZ} boundaries

6.2 Stage 2: Rule Validation

1. **Below C_{ground} :** Verify that dynamics preserve information measures (e.g., predictability, reversibility)
2. **Above C_{ground} :** Detect signatures of information reduction (e.g., entropy increase, memory loss)
3. **Transition:** Study critical behavior near the boundary

6.3 Stage 3: Robustness Assessment

1. **Proxy Variation:** Test with alternate complexity measures (statistical complexity, approximate entropy)
2. **Encoding Sensitivity:** Vary encoding schemes within physically reasonable bounds
3. **Scale Independence:** Verify boundary stability across system sizes

Example 6.1 (Application to Quantum-Classical Transition). *One might encode quantum states via their Wigner function discretizations, compute K_{LZ} on time series of these encodings, and test whether the quantum-classical transition correlates with crossing a K_{LZ} threshold C_{ground} . The framework provides the formal structure to ensure such investigations maintain logical consistency.*

7 Relation to Prior Work

Our approach complements but does not subsume existing information-centric programs:

- **Quantum Darwinism [3]**: We provide formal tools to test hypotheses about branching structures
- **Emergent Gravity [5]**: Our framework offers rigorous complexity notions for holographic scenarios
- **Computational Universe [4]**: We add formal verification to complexity-based physical models

Unlike these programs, we focus on providing a *methodological foundation* with machine-checked logical guarantees.

8 Formal Verification and Artifact Availability

All definitions and theorems are implemented and verified in Lean 4. Key verification achievements include:

- Complete mechanization of the three-layer architecture
- Machine-checked proofs of all bridge theorems
- Zero outstanding proof gaps (`sorry` declarations)

To support reproducibility:

- Versioned, anonymized repository provided as supplementary material
- DOI archive will be released upon acceptance

9 Limitations and Future Directions

- **Empirical Calibration**: C_{ground} must be determined experimentally for each domain; no universal value is claimed
- **Proxy Dependence**: Conclusions should be robust across different complexity measures
- **Scale Limitations**: The framework currently addresses finite, discrete systems
- **Physical Identification**: Specific identifications (e.g., “ R_{red} equals wavefunction collapse”) remain hypotheses

Future work will develop concrete instantiations for quantum measurement, cosmological structure formation, and neural computation.

10 Conclusion

Substrate Theory provides a formally verified framework for studying complexity-threshold dynamics with clear separation between ideal concepts and operational measurements. By treating the regime boundary as an empirical parameter and distinguishing mathematical theorems from physical postulates, we enable disciplined, testable modeling of informational regimes in physical systems. The machine-checked formalization ensures logical consistency, while the explicit empirical interface enables meaningful confrontation with experimental data.

Acknowledgments We thank colleagues for discussions on algorithmic information, mechanized reasoning, and model validation. All remaining errors are our own.

References

- [1] A. N. Kolmogorov. Three approaches to the quantitative definition of information. *Problems of Information Transmission*, 1(1):1–7, 1965.
- [2] J. Ziv and A. Lempel. A universal algorithm for sequential data compression. *IEEE Transactions on Information Theory*, 23(3):337–343, 1977.
- [3] W. H. Zurek. Decoherence, einselection, and the quantum origins of the classical. *Reviews of Modern Physics*, 75(3):715–775, 2003.
- [4] S. Lloyd. Computational capacity of the universe. *Physical Review Letters*, 88(23):237901, 2002.
- [5] E. Verlinde. On the origin of gravity and the laws of Newton. *Journal of High Energy Physics*, 2011(4):29, 2011.

A Lean 4 Formalization Excerpts

This appendix contains selected Lean code snippets illustrating the formalization. Complete source is available in the supplementary material.

A.1 Bridge Layer Theorems

```
-- Bridge between ideal and operational complexity
theorem proxy_calibration :
  ∃ a b : ℝ, a ≥ 0 ∧ b ≥ 0 ∧
  ∀ (x y : State), KLZ (x ++ y) ≤ KLZ x + KLZ y + a := by
  ...
```

A.2 Operational Rules

```
-- Coherence rule application
theorem coherence_preservation :
  ∀ (n h : List State), coherent_state (join n) →
  K_LZ (R_Cohesion n h) = K_LZ h := by
  ...
```