Discrete Causal Topology

Victor Zhorin

December 2025

Abstract

We prove that continuous dynamics emerge from discrete causal structure through a mechanism we call *localization*: the extraction of where a system sits in configuration space rather than enumeration of transitions between discrete states. Three necessity theorems establish that optimal prediction from discrete event sequences with terminal boundaries requires preservation of event identity, spectral temporal encoding, and learned causal attention. Each component is individually irreducible. The architecture is not engineered but derived from information-theoretic constraints.

The central result concerns gradient flow from boundary conditions. We prove that terminal outcomes shape the entire antecedent causal field through backpropagation, creating second-order structure where early events modify interpretation of later events. This retroactive determination parallels the holographic principle: predictive information satisfies an area law scaling with event count, not temporal duration.

The framework unifies phenomena across substrates because localization creates shared coordinates that enable cross-trajectory learning: populations of event sequences reveal the universal geometry of progression toward boundary. Any system generating discrete timestamped events with terminal boundaries instantiates the same mathematical structure, dissolving the distinction between exact and statistical sciences. We prove why transformer architectures succeed and establish that positional encoding is categorically incorrect for timestamped data. Empirical predictions are falsifiable: architectures satisfying the necessity theorems should extract continuous dynamics that tabular methods systematically discard.

Contents

1 Introduction

How does continuous description emerge from discrete structure?

This question appears wherever physics confronts foundations. Quantum mechanics delivers discrete spectra; thermodynamics requires continuous variables. Particle interactions are countable events; scattering amplitudes are continuous functions. Causal set theory proposes discrete spacetime events; general relativity assumes smooth manifolds.

The standard resolution discretizes the continuum: lattice approximations, cutoff regularization, finite element methods. But this inverts the ontology. If discrete structure is fundamental, continuous description should emerge, not be imposed and subsequently approximated.

We establish emergence through *localization*: continuous observables arise from determining where a system sits in configuration space, not from counting transitions between discrete states. The mechanism operates identically across domains.

1.1 The Discrete-Continuous Tension

Consider three manifestations:

Spectroscopy. Ions in crystal lattices occupy discrete sites. Optical transitions occur at discrete frequencies. Yet absorption spectra are continuous curves. Standard theory sums over vibrational quanta. The continuous spectrum emerges as infinite series over discrete transitions.

Prediction. Events in time are discrete occurrences. Yet optimal prediction requires modeling continuous risk dynamics. Standard approaches fit continuous hazard functions or compress sequences through recurrent hidden states simulating continuous evolution.

Spacetime. Causal set theory proposes that spacetime events form a locally finite poset, with continuous Lorentzian geometry emerging through coarse-graining. The approach is mathematically coherent but empirically inaccessible at Planck scale.

In each case, discrete structure is primary; continuous description is derived. We prove this derivation follows a universal pattern.

1.2 Main Results

1. **Localization Principle.** Continuous observables emerge from detecting system position in configuration space without enumerating discrete transitions (Section 2).

2. **Necessity Theorems.** Optimal extraction of continuous dynamics from discrete event sequences requires three irreducible components: discrete event representation, spectral temporal encoding, and learned causal attention (Section 4).

3. **Boundary Determination.** Terminal conditions shape the entire antecedent field through gradient flow, creating second-order causal structure (Section 5).

4. **Holographic Information Bound.** Predictive information satisfies an area law: it scales with event count (boundary), not temporal duration (bulk) (Section 6).

5. **Substrate Independence.** Localization creates shared coordinates enabling cross-trajectory learning. Populations of event sequences reveal universal geometry of progression toward boundary. The mathematical structure applies uniformly across physical, biological, and social systems (Section 7).

2 Localization

2.1 The Principle

A measurement reveals where a system is, not how it got there.

This observation, elementary in classical mechanics, has non-trivial implications for emergence. When we ask how continuous observables arise from discrete structure, the standard answer traces paths through discrete state space. Localization offers an alternative: the observable is determined by instantaneous configuration; the path is irrelevant.

Definition 2.1 (Configuration Space)**.** *A configuration space $\mathcal{C}$ is a set equipped with topology. A configuration is a point $q \in \mathcal{C}$. An observable is a continuous function $O : \mathcal{C} \to \mathbb{R}$.*

Definition 2.2 (Localization)**.** *An observation $O(q)$ is obtained by localization if it depends only on the instantaneous configuration q, not on the trajectory $q(t)$ leading to it.*

2.2 Spatial Configuration

In optical spectroscopy, the configuration space is the lattice potential landscape. An ion's absorption probability at frequency ω depends on where it sits in this landscape at the moment of photon interaction:

$$W(\omega; q) = \frac{2\pi}{\hbar} |\langle f|\hat{\mu}(q)|i\rangle|^2 \delta(\omega - \omega_{if}(q)) \tag{1}$$

The continuous spectrum emerges not from summing over phonon states but from averaging over configuration space:

$$I(\omega) = \int_{\mathcal{C}} W(\omega; q)\, \rho(q)\, dq \tag{2}$$

where $\rho(q)$ is the thermal distribution over configurations.

No phonon counting. The light detects where the ion sits. The continuum emerges from the probability distribution over positions.

2.3 Temporal Configuration

Extend the principle to time. An event's predictive content depends on where it sits in causal configuration.

Definition 2.3 (Causal Configuration)**.** *For event e in event space $\mathcal{M}$, its causal configuration is the pair $(J^-(e), \tau(e))$ where:*

- *$J^-(e) = \{e' \in E : e' \prec e\}$ is the causal past*
- *$\tau(e) \in \mathbb{R}^+$ is the temporal coordinate*

Just as an ion's spectral signature depends on spatial position in the lattice potential, an event's predictive contribution depends on temporal position in causal structure.

The continuous dynamics of risk, trajectory, evolution: these emerge from the distribution over causal configurations, not from counting transitions between discrete states.

3 Event Topology

3.1 Formal Structure

Definition 3.1 (Event Space)**.** *An event space is a tuple $\mathcal{M} = (E, \prec, \tau, \kappa)$ where:*

- *E is a finite set of events*
- *$\prec$ is a partial order on E (causal precedence)*
- *$\tau : E \to \mathbb{R}^+$ is order-preserving: $e \prec e' \Rightarrow \tau(e) < \tau(e')$*
- *$\kappa : E \to C$ assigns event types from finite vocabulary C*

Definition 3.2 (Causal Structure). *The causal past is* $J^-(e) = \{e' : e' \prec e\}$. *The causal future is* $J^+(e) = \{e' : e \prec e'\}$. *The causal diamond is* $\Diamond(e_i, e_j) = J^+(e_i) \cap J^-(e_j)$.

Definition 3.3 (Boundary). *An event* $e^\dagger$ *is a terminal boundary if* $J^+(e^\dagger) = \emptyset$. *A prediction horizon* $T < \tau(e^\dagger)$ *defines the observable past* $J^-(T) = \{e : \tau(e) < T\}$.

3.2 The Prediction Problem

Given event space $\mathcal{M}$ observed up to horizon T, predict the terminal boundary event $e^\dagger$.

Definition 3.4 (Predictor). *A predictor is a measurable function* $\mu : \mathcal{M}|_T \to [0, 1]$ *where* $\mathcal{M}|_T$ *denotes the event space restricted to* $J^-(T)$.

The problem: characterize optimal predictors μ^* minimizing expected loss over distribution $p(\mathcal{M})$.

4 Necessity Theorems

We prove that optimal prediction requires three architectural components, each individually irreducible.

4.1 Information Bound

Axiom 4.1 (Holographic Bound). *Predictive mutual information satisfies:*

$$I(\mu(\mathcal{M}); e^\dagger) \leq C \cdot |J^-(T)| \tag{3}$$

where C *is bits per event and* $|J^-(T)|$ *is the event count in observable past.*

Remark 4.2. *Information scales with event count (the boundary of causal past), not temporal duration (the bulk). This is the holographic principle: bulk information is bounded by boundary area.*

4.2 Necessity of Discrete Event Representation

Theorem 4.3 (Discrete Events are Necessary). *Let* $f : \mathcal{P}(C) \times \mathcal{P}(\mathbb{R}^+) \to [0, 1]$ *depend only on marginal distributions of event types and timestamps. There exist event spaces* $\mathcal{M}_1, \mathcal{M}_2$ *satisfying:*

1. *Identical type marginals:* $\pi_C(\mathcal{M}_1) = \pi_C(\mathcal{M}_2)$
2. *Identical time marginals:* $\pi_\tau(\mathcal{M}_1) = \pi_\tau(\mathcal{M}_2)$
3. *Different optimal predictions:* $\mu^*(\mathcal{M}_1) \neq \mu^*(\mathcal{M}_2)$

Thus $f(\mathcal{M}_1) = f(\mathcal{M}_2)$ *but the prediction task requires distinguishing them.*

Proof. Construct:

$$\mathcal{M}_1 = \{(A, t_1) \prec (B, t_2)\} \tag{4}$$

$$\mathcal{M}_2 = \{(B, t_1) \prec (A, t_2)\} \tag{5}$$

Both have type set $\{A, B\}$ and timestamp set $\{t_1, t_2\}$, hence identical marginals. Any function of marginals gives $f(\mathcal{M}_1) = f(\mathcal{M}_2)$.

However, causal order carries predictive information. If $A \prec B$ indicates causal cascade while $B \prec A$ indicates independent occurrence, optimal predictions differ.

Formally: let $p(e^\dagger = 1 | A \prec B) = 0.8$ and $p(e^\dagger = 1 | B \prec A) = 0.3$. Then $\mu^*(\mathcal{M}_1) = 0.8 \neq 0.3 = \mu^*(\mathcal{M}_2)$, but any marginal-based function cannot distinguish them. □

Corollary 4.4. *Tabular representations aggregating event counts discard predictive information encoded in causal order.*

4.3 Necessity of Spectral Temporal Encoding

Definition 4.5 (Ordinal Encoding)**.** *An encoding $\phi : E \to \mathbb{R}^d$ is ordinal if it depends only on sequence position: $\phi(e_k) = PE(k)$ for some fixed function $PE : \mathbb{N} \to \mathbb{R}^d$.*

Theorem 4.6 (Temporal Encoding is Necessary)**.** *Let $g : C^n \to [0, 1]$ depend only on the sequence of event types with ordinal encoding. There exist event spaces with identical type sequences but different temporal geometry requiring different optimal predictions.*

Proof. Construct:

$$\mathcal{M}_1 = \{(c, 0), (c, 1), (c, 2)\} \tag{6}$$
$$\mathcal{M}_2 = \{(c, 0), (c, \epsilon), (c, T)\} \tag{7}$$

for small ϵ and large T.

Both have type sequence (c, c, c). Ordinal encoding assigns positions $(1, 2, 3)$ to both. Thus $g(\mathcal{M}_1) = g(\mathcal{M}_2)$.

However, $\mathcal{M}_1$ has uniform inter-event spacing (stable rate), while $\mathcal{M}_2$ has rapid initial events then quiescence (burst then silence). These represent geometrically distinct trajectories.

Let $p(e^\dagger = 1|\text{uniform}) = 0.2$ and $p(e^\dagger = 1|\text{burst}) = 0.7$. Optimal predictions differ; ordinal encoding cannot distinguish them. □

Definition 4.7 (Spectral Temporal Encoding)**.** *A spectral temporal encoding is:*

$$\Phi_\theta(t) = (\omega_0 t + \phi_0, \sin(\omega_1 t + \phi_1), \ldots, \sin(\omega_{d-1} t + \phi_{d-1})) \tag{8}$$

with learnable parameters $\theta = \{\omega_k, \phi_k\}$.

Theorem 4.8 (Universal Approximation)**.** *For any $f \in C([0, T])$ and $\varepsilon > 0$, there exist θ and linear $L : \mathbb{R}^d \to \mathbb{R}$ with $\|f - L \circ \Phi_\theta\|_\infty < \varepsilon$.*

Proof. The set $\mathcal{A} = \{1, t, \sin(\omega t + \phi) : \omega, \phi \in \mathbb{R}\}$ generates a subalgebra of $C([0, T])$ that separates points (via the linear function t) and vanishes nowhere (contains constant 1). By Stone-Weierstrass, $\mathcal{A}$ is uniformly dense. Any $f \in C([0, T])$ is approximable by linear combinations from $\mathcal{A}$, which is precisely $L \circ \Phi_\theta$ for appropriate θ. □

Remark 4.9. *The frequencies ω_k are discovered by gradient descent, not prescribed. Learning identifies characteristic timescales at which causal structure operates.*

4.4 Necessity of Causal Attention

Definition 4.10 (Fixed Aggregation)**.** *A fixed aggregation predictor has form:*

$$h(\mathcal{M}) = F\left(\sum_{e \in E} w(\kappa(e)) \cdot \psi(e)\right) \tag{9}$$

where weights $w : C \to \mathbb{R}$ depend only on event type, not context.

Theorem 4.11 (Attention is Necessary)**.** *There exist prediction tasks where optimal prediction requires context-dependent event interpretation that fixed aggregation cannot provide.*

Proof. Let event type c appear in two contexts:

$$\mathcal{M}_A = \{(c_1, t_1) \prec (c, t_2)\} \tag{10}$$
$$\mathcal{M}_B = \{(c_2, t_1) \prec (c, t_2)\} \tag{11}$$

where c_1, c_2 are background conditions that modulate interpretation of c.

Fixed aggregation assigns weight $w(c)$ regardless of context. But if c_1 indicates c is dangerous while c_2 indicates c is benign:

$$p(e^\dagger = 1 | c_1 \prec c) = 0.9 \tag{12}$$

$$p(e^\dagger = 1 | c_2 \prec c) = 0.1 \tag{13}$$

Optimal prediction requires context-dependent weights: $\alpha(c|c_1) \neq \alpha(c|c_2)$.

Define causal attention:

$$\alpha(e_i, e_j) = \frac{\exp\langle Q\psi(e_i), K\psi(e_j)\rangle}{\sum_{e_k \prec e_i} \exp\langle Q\psi(e_i), K\psi(e_k)\rangle} \tag{14}$$

This provides context-dependent weighting that fixed aggregation cannot. □

4.5 The Minimal Architecture

Theorem 4.12 (Minimal Causal Encoding)**.** *Under Axiom 4.1, optimal predictors admit factorization:*

$$\mu^*(\mathcal{M}) = F \circ \mathcal{T}_\alpha^L \circ (\psi \oplus \Phi_\theta \circ \tau) \tag{15}$$

where:

- $\psi : C \to \mathbb{R}^d$ *embeds event types (Theorem 4.3)*
- $\Phi_\theta \circ \tau$ *spectrally encodes timestamps (Theorem 4.6)*
- $\mathcal{T}_\alpha^L$ *applies L layers of causal attention (Theorem 4.11)*
- $F : \mathbb{R}^d \to [0, 1]$ *is output transformation*

Each component is necessary; removing any one reduces achievable performance.

Proof. Existence: The class of functions with this form is non-empty and closed under limits in appropriate topology. Under compact parameter constraints, optimal solutions exist by Weierstrass.

Necessity: Theorems 4.3, 4.6, 4.11 prove each component is independently required.

Uniqueness: Up to gauge transformations $\psi \mapsto O\psi$, $Q \mapsto QO^T$, $K \mapsto KO^T$ for orthogonal O, which preserve attention scores, the factorization is unique. □

5 Gradient Flow and Boundary Determination

The necessity theorems establish architectural requirements. We now prove the deeper result: terminal boundary conditions determine the entire antecedent causal field through gradient flow.

5.1 The Boundary Principle

In general relativity, boundary conditions (initial data on a spacelike surface) determine the entire spacetime evolution. In thermodynamics, boundary conditions (temperature at surfaces) determine equilibrium distributions throughout the bulk.

We prove an analogous principle: the terminal boundary event $e^\dagger$ determines the learned causal weights throughout the antecedent event structure.

5.2 Gradient Decomposition

Theorem 5.1 (Path Decomposition of Gradients). *Let $\mathcal{L}(\theta) = -\log p(e^\dagger|\mathcal{M};\theta)$ be the prediction loss. The gradient with respect to event embedding $\psi(e_j)$ decomposes over causal paths:*

$$\frac{\partial \mathcal{L}}{\partial \psi(e_j)} = \sum_{\gamma \in \Gamma(e_j, e^\dagger)} w_\gamma \cdot v_\gamma \tag{16}$$

where $\Gamma(e_j, e^\dagger)$ is the set of directed paths from e_j to boundary, w_γ is path weight, and v_γ is path direction.

Proof. By chain rule through L attention layers:

$$\frac{\partial \mathcal{L}}{\partial \psi(e_j)} = \sum_{\ell=1}^{L} \frac{\partial \mathcal{L}}{\partial h^{(\ell)}} \cdot \frac{\partial h^{(\ell)}}{\partial \psi(e_j)} \tag{17}$$

Each layer propagates through attention:

$$h_i^{(\ell)} = \sum_{k: e_k \prec e_i} \alpha^{(\ell)}(e_i, e_k) \cdot V^{(\ell)} h_k^{(\ell-1)} \tag{18}$$

The contribution of $\psi(e_j)$ to $h^{(L)}$ flows through all paths $\gamma : e_j \prec \cdots \prec e_i$ where e_i influences the output. Each path contributes weight:

$$w_\gamma = \prod_{(e_{k'}, e_{k''}) \in \gamma} \alpha(e_{k'}, e_{k''}) \tag{19}$$

Summing over all paths gives the stated decomposition. □

5.3 Retroactive Causation

Definition 5.2 (Causal Susceptibility). *The causal susceptibility tensor is:*

$$\chi(e_i, e_j) = \frac{\partial^2 \mathcal{L}}{\partial \psi(e_i) \partial \psi(e_j)} \tag{20}$$

Theorem 5.3 (Second-Order Structure). *The susceptibility $\chi(e_i, e_j)$ is generically non-zero for causally related events, capturing how e_i modifies the predictive impact of e_j.*

Proof. For $e_i \prec e_j$, the path from e_i to boundary passes through e_j. The attention weight $\alpha(e_j, e_i)$ depends on both $\psi(e_i)$ and $\psi(e_j)$ through the softmax:

$$\alpha(e_j, e_i) = \frac{\exp\langle Q\psi(e_j), K\psi(e_i)\rangle}{\sum_{k \prec j} \exp\langle Q\psi(e_j), K\psi(e_k)\rangle} \tag{21}$$

Thus:

$$\frac{\partial \alpha(e_j, e_i)}{\partial \psi(e_i)} \neq 0, \quad \frac{\partial \alpha(e_j, e_i)}{\partial \psi(e_j)} \neq 0 \tag{22}$$

The mixed partial $\partial^2 \mathcal{L} / \partial \psi(e_i) \partial \psi(e_j)$ inherits non-zero contributions from the chain rule. □

Remark 5.4. *This is retroactive determination: event e_i occurring early in the sequence modifies how later event e_j is interpreted. The modification is not forward causation (events cannot influence their past) but learned correlation structure shaped by the terminal boundary through gradient descent.*

5.4 The Boundary Determines the Field

Corollary 5.5 (Boundary Determination Principle)**.** *At convergence, the learned attention weights α^* satisfy:*

$$\alpha^*(e_i, e_j) \propto \frac{\partial}{\partial \psi(e_j)} \log p(e^\dagger | \mathcal{M}) \tag{23}$$

The terminal boundary determines the entire causal weighting structure.

Proof. At critical points of $\mathcal{L}$, gradients vanish: $\partial \mathcal{L}/\partial \theta = 0$. The attention weights are fixed points of gradient flow from the boundary. Different boundaries (different terminal outcomes in training distribution) yield different weight configurations.

This parallels general relativity: boundary data on a surface determines the geometry throughout the bulk. Here, boundary data (terminal events) determines the causal geometry (attention weights) throughout the antecedent structure. □

6 Holographic Structure

6.1 The Bekenstein Analogy

Bekenstein (1973) proved that black hole entropy is proportional to horizon area, not enclosed volume:

$$S_{BH} = \frac{k_B c^3}{4G\hbar} A \tag{24}$$

This led to the holographic principle: information in a region is bounded by its boundary area.

6.2 Information Bound in Event Topology

Theorem 6.1 (Area Law for Predictive Information)**.** *Under Axiom 4.1, predictive information scales with event count:*

$$I(\mu; e^\dagger) \leq C \cdot |J^-(T)| \tag{25}$$

This is an area law: information scales with the boundary of causal past (event count), not with bulk volume (temporal duration).

Proof. Each event carries at most C bits of predictive information (finite vocabulary, bounded correlations). Events are the atoms of the causal structure; their count is the natural measure of boundary size.

Temporal duration is the bulk: a period of 10 years with 5 events contains less information than a period of 1 year with 100 events. Information is where events are, not where time passes. □

6.3 The CLS Token as Holographic Screen

In the architecture of Theorem 4.12, a classification token aggregates information from all events through attention:

$$h_{\text{CLS}} = \sum_{e \in E} \alpha(\text{CLS}, e) \cdot V\psi(e) \tag{26}$$

Proposition 6.2 (Holographic Encoding)**.** *The CLS representation h_{CLS} encodes all predictive information from the event sequence, with attention weights $\alpha(CLS, e)$ acting as the holographic dictionary mapping bulk (events) to boundary (representation).*

Proof. By Theorem 4.12, optimal prediction factors through h_{CLS}. By Theorem 6.1, this representation contains all extractable predictive information, bounded by event count. The attention weights determine which bulk events contribute to the boundary representation. □

6.4 Correspondence Table

Holography (Spacetime)	Event Topology
Black hole interior	Causal past $J^-(T)$
Event horizon	Prediction horizon T
Horizon area	Event count $\lvert J^-(T)\rvert$
Bekenstein entropy $S \propto A$	Information bound $I \leq C\lvert J^-(T)\rvert$
Holographic screen	CLS token
Bulk-boundary dictionary	Attention weights α
Boundary determines bulk	Gradient flow from $e^\dagger$

7 Substrate Independence

7.1 Why Universality Holds

The necessity theorems establish what architecture is required. We now explain *why* the same mathematics governs disparate substrates.

The key insight: localization in learned coordinates enables cross-trajectory interrogation.

Consider a population of trajectories $\{\mathcal{M}_1, \mathcal{M}_2, \ldots, \mathcal{M}_N\}$, each a sequence of discrete events terminating at boundary. Spectral temporal encoding places events from all trajectories into a shared coordinate system:

$$\Phi_\theta : \mathbb{R}^+ \to \mathbb{R}^d \tag{27}$$

Events at similar temporal coordinates, across different trajectories, become comparable. The attention weights learned from the population encode:

$$\alpha^*(e_i, e_j) = f(\text{“event type } \kappa(e_j) \text{ at temporal location } \Phi_\theta(\tau(e_j)) \text{ predicts boundary”}) \tag{28}$$

This is population-level knowledge compiled into individual prediction. The learned frequencies $\{\omega_k\}$ are not arbitrary basis functions but the characteristic timescales at which causal structure operates across the population.

Theorem 7.1 (Cross-Trajectory Learning)**.** *The continuum extracted by localization is not the individual trajectory smoothed. It is the population-learned manifold of progression toward boundary.*

Proof. Gradient flow from boundaries (Section 5) shapes attention weights. The loss aggregates over the population:

$$\mathcal{L}(\theta) = \sum_{i=1}^{N} \mathcal{L}_i(\theta) = -\sum_{i=1}^{N} \log p(e_i^\dagger | \mathcal{M}_i; \theta) \tag{29}$$

The learned parameters θ^* minimize population loss, not individual loss. The spectral frequencies, attention weights, and embedding geometry encode structure shared across trajectories. Individual prediction applies this population-learned structure to locate a single trajectory within the shared manifold. □

7.2 The Universal Geometry

What is this shared manifold? It is the phase space of evolution toward termination.

Stars traverse it (nucleosynthesis, shell burning, collapse). Patients traverse it (health, disease, deterioration). Companies traverse it (growth, stress, failure). Networks traverse it (normal operation, anomaly, breach).

The substrates differ. The geometry is identical: a space of causal configurations where discrete events mark positions and terminal boundaries define the attractor.

Theorem 7.2 (Universality)**.** *Any system satisfying the event space axioms (Definition 3.1) with terminal boundary admits the minimal architecture of Theorem 4.12, because:*

1. *Discrete events require preservation (Theorem 4.3)*
2. *Temporal geometry requires spectral encoding (Theorem 4.6)*
3. *Context-dependence requires attention (Theorem 4.11)*
4. *Population learning requires shared coordinates (Theorem 7.1)*

The architecture is not designed but derived from the structure of the prediction problem itself.

7.3 Instances

Domain	Events	Boundary	Shared Manifold
Stellar evolution	Nucleosynthesis stages	Supernova/collapse	Hertzsprung-Russell trajectory
Medicine	Encounters	Death	Disease progression space
Finance	Transactions	Default	Credit deterioration manifold
Networks	Packets	Breach	Threat evolution surface
Particle physics	Interactions	Decay	Amplitude landscape
Spacetime	Events	Singularity	Lorentzian geometry

In each case, the continuum is not imposed but discovered: the population of trajectories reveals the geometry of progression toward boundary, and localization places individual trajectories within this learned structure.

7.4 Dissolution of Science Hierarchy

The distinction between "hard" sciences (physics, chemistry) and "soft" sciences (psychology, economics, sociology) rests on an assumption: physical systems have mathematical structure while social systems have only statistical regularities.

Corollary 7.3 (Dissolution)**.** *Any system generating discrete timestamped events with terminal boundaries satisfies identical mathematical constraints. The hard/soft distinction is artifact, not ontology.*

Proof. The mathematics depends on three properties: discrete events, temporal coordinates, terminal boundary. These are structural, not material. A stellar trajectory and a patient trajectory instantiate the same causal topology because both are populations of event sequences terminating at boundaries. Localization creates shared coordinates. Shared coordinates enable cross-trajectory learning. Cross-trajectory learning discovers universal geometry. The substrate is irrelevant; the structure is everything. □

8 Implications for Machine Learning

8.1 Why Transformers Work

The transformer architecture (Vaswani et al., 2017) was discovered empirically. Eight years later, theoretical foundation remains absent.

Theorem 8.1 (Transformer Necessity)**.** *For prediction tasks on discrete event sequences with terminal boundaries, transformer-like architectures with spectral temporal encoding are necessary for optimal performance.*

Proof. Direct consequence of Theorems 4.3, 4.6, 4.11. Self-attention implements the causal influence kernel. Multi-head attention discovers multiple causal channels. Layer composition builds hierarchical abstraction. These are not design choices but mathematical requirements. □

8.2 Positional Encoding is a Category Error

"Positional encoding" encodes ordinal index (1st, 2nd, 3rd), not position. The name is a misnomer that obscures a mathematical obstruction.

For language, ordinal index is all that exists. Tokens have no timestamps. $\text{PE}(k) = \sin(k/10000^{2i/d})$ assigns indices to sequence slots. This is appropriate: "the cat sat" differs from "sat the cat" only in order.

For timestamped events, ordinal index discards the actual temporal geometry. Two sequences with identical event types in identical order but different temporal spacing require different predictions. Theorem 4.6 proves this is not a limitation but a mathematical obstruction: ordinal encoding provably loses information that affects optimal prediction.

Corollary 8.2 (Categorical Error)**.** *For timestamped event sequences, ordinal positional encoding is provably suboptimal. Spectral temporal encoding is necessary.*

Proof. Theorem 4.6. Ordinal encoding discards temporal geometry that affects optimal prediction. □

Spectral temporal encoding performs actual localization: $\Phi_\theta(t)$ extracts where events sit in continuous time, learning characteristic frequencies at which causal structure operates. This is localization applied to time.

Transformers succeeded on language first. Language has no time. The world does. Extension to physical, biological, and social systems requires replacing the misnomer with localization.

9 Empirical Predictions

9.1 Falsifiability

The framework generates testable predictions.

Prediction 9.1 (Architecture Sufficiency)**.** *Systems satisfying Theorem 4.12 will extract continuous dynamics that tabular methods and ordinal-encoded transformers cannot access.*

Prediction 9.2 (Performance Gap Closure)**.** *In domains where data-rich and data-sparse settings show performance gaps attributed to missing information, architectures satisfying the necessity theorems will close the gap, revealing the gap as architectural, not informational.*

9.2 Why Medicine?

Medical prediction is optimal for validation:

- Events are discrete, timestamped (encounters have dates)
- Causal ordering is unambiguous (time flows forward)
- Boundaries are binary, verifiable (mortality is observed)
- Datasets are large (millions of trajectories)

- Comparisons exist (data-rich vs. data-sparse settings)

Success establishes the mathematics. Substrate independence (Theorem 7.2) guarantees transfer to other domains.

10 Discussion

10.1 Summary of Claims

1. **Localization** is the mechanism of discrete-to-continuous emergence: observables reveal where systems sit in configuration space.
2. **Three components are necessary**: discrete events, spectral temporal encoding, and causal attention. Each is irreducible.
3. **Boundaries determine fields**: terminal conditions shape antecedent causal structure through gradient flow.
4. **Information satisfies an area law**: predictive content scales with event count, not temporal duration.
5. **Substrate independence follows from shared coordinates**: localization enables cross-trajectory learning; populations reveal universal geometry; physical, biological, and social systems instantiate the same causal topology because structure, not substrate, determines the mathematics.

10.2 What This Does Not Claim

- That current implementations are optimal (they approximate the necessary structure)
- That all prediction tasks require this architecture (only those with discrete timestamped events and terminal boundaries)
- That social systems reduce to physical systems (both instantiate common mathematical structure)

10.3 The Emergence Program

Physics has sought to derive continuous description from discrete structure since quantum mechanics revealed discreteness at fundamental scales. Causal set theory proposes discrete spacetime but lacks empirical access at Planck scale.

Event topology provides a laboratory. The mathematics of discrete-to-continuous emergence operates at human scales with abundant data. The necessity theorems are provable; the predictions are testable.

Whether the same mathematical structure governs Planck-scale physics is conjecture. That it governs human-scale prediction is theorem.

The oscillator sits in configuration space.
The event sits in causal time.
Localization extracts the continuum.

References

[1] J.D. Bekenstein, "Black holes and entropy," *Phys. Rev. D* **7**, 2333 (1973).

[2] L. Bombelli, J. Lee, D. Meyer, R.D. Sorkin, "Space-time as a causal set," *Phys. Rev. Lett.* **59**, 521 (1987).

[3] R.D. Sorkin, "Causal sets: Discrete gravity," in *Lectures on Quantum Gravity*, Springer (2003).

[4] L. Susskind, "The world as a hologram," *J. Math. Phys.* **36**, 6377 (1995).

[5] A. Vaswani et al., "Attention is all you need," *Advances in Neural Information Processing Systems* **30** (2017).