

Genomic Sequences as Discrete Spacetime Condensates: An End-to-End Generative Physics Framework via Gauge Symmetry Breaking

Wanhong Huang
Serendip Commons Society
huangwanhong@serendip.ngo

Abstract

We investigate an end-to-end generative-physics hypothesis for genomic structure with explicit separation of verified finite models and the proposed microscopic and condensate mechanisms. The constructive foundation is an oriented cellular two-complex with a designated genomic boundary and a normalized four-state sequence/contact model. We prove its cellular incidence identities, exact local $\mathbb{Z}_2$ gauge invariance, and the gauge quotient that reduces the boundary theory to four observable states. Binary contact variables can be marginalized exactly, yielding a pairwise Potts-type law and exposing a sequence-only parameter-identifiability obstruction. Small-system exact enumeration and exploratory Python sampling check these identities; these tests establish computational consistency at the scale of the constructed finite systems; biological benchmarking remains a separate objective. A separate smooth-field discretization has a controlled second-order continuum error, while the uncoupled sequence model obeys an exact block-composition fluctuation law. Both results provide limited mathematical benchmarks; an interacting continuum limit and a physical condensate require further derivation. Nonlinear nucleotide readout also fails, in general, to commute with coarse-graining.

Independently of any microscopic group-field derivation, we develop data-informed genomic two-complex analysis through persistent and relative homology, cellular maps, attributed cochains, and optional sheaf or symmetry structures. We specify the face-construction assumptions required for these invariants to be biologically interpretable. Fixed-complex homology is invariant under changes to nucleotide labels; truncating higher-dimensional complexes and changing resolution may alter measured topology. We derive conditional stability and correspondence results, including bounds for uncertain contact scores, multiresolution chain maps, partial alignments, and probability distributions over alternative genome correspondences. Finite stochastic examples illustrate identifiability limitations, intervention-dependent distinctions, topology-changing histories, and Bayesian filtering of latent cellular events. A genome-scale implementation architecture and benchmarks are proposed, not claimed to have been realized.

Finally, an information-theoretic analysis shows why observation maps preserve only distinctions retained by their measurement channels and why deterministic topological descriptors preserve the information ceiling set by the full input available to an ideal predictor. The framework's prospective scientific value therefore lies in demonstrable additional evidence, interpretable constraints, improved inference, or independently validated physical predictions. Derivation of a genomic cellular-history law from a microscopic group-field action, chemically grounded nucleotide formation, a gauge-invariant condensate phase, and a nontrivial geometric continuum remain open research problems.

Keywords: genomic generativity; relational complexes; lattice gauge invariance; genomic two-complexes; persistent homology; cellular sheaves; comparative genomes; group-field theory; coarse-graining; continuum limits.

License: © 2026 Wanhong Huang. This manuscript is distributed under the [Creative Commons Attribution 4.0 International \(CC BY 4.0\)](https://creativecommons.org/licenses/by/4.0/) license.

1 Introduction

The common representation of a genome is an ordered word $w \in \Sigma^N$, with $\Sigma = \{A, C, G, T\}$. This is an indispensable observable and computational description, yet by itself the additional questions of polymer production and rearrangement, nonlocal physical interactions, and large-scale structural emergence call for a complementary generative description. The preceding Group-Equivariant Context-Sensitive Grammar (GE-CSG) proposal (Huang, 2026) explored a group-algebraic route to deriving symbolic structure instead of declaring a nucleotide alphabet to be an uninterpreted primitive. Its proposal to connect group representations, symmetry reduction, and context-sensitive derivations motivates a more explicitly physical investigation.

A physical generative theory requires an energy function or kinetic generator, admissible transformations, and a model of history in addition to formal symmetry. Moreover, a grammatical rewriting relation may include irreversible transformations and can therefore require algebraic structures beyond group actions. A theory of generation must say what the admissible configurations are; which microscopic operations alter them; what constraints and weights govern these operations; and which observations make them distinguishable. We propose to treat sequence formation as a *readout of a generative relational structure*, while recognizing that the actual chemical identities of nucleotides require molecular parameters external to the abstract toy model.

An analogy with group-field theory (GFT) is suggestive but delicate. Interacting group-valued fields can generate combinatorial complexes, and particular collective states have been investigated as candidates for effective continuous geometry in quantum gravity (Orti, 2014; Gielen et al., 2013, 2014). Their application to genomic matter remains a hypothesis that requires separate physical justification. In chromatin physics, by contrast, polymer models with state-dependent interactions and nonequilibrium activities are already biologically motivated (Michieletto et al., 2016; Forte et al., 2026). Recent work also reviews the empirical and computational basis for a sequence grammar of three-dimensional genome organisation (Tamon et al., 2025). Our aim is to specify a mathematical bridge while keeping its status clear.

Well-developed alternatives make the required incremental contribution demanding: genealogy inference already represents population histories, pangenome graphs capture variation, chromatin polymer models compute physical configurations, and recent sequence models predict regulatory consequences at megabase context. Sections 30–31 therefore formulate distinct, falsifiable contributions rather than treating a new representation as a performance result.

We make five contributions. First, we construct an explicit cellular history support with incidence operators, an oriented terminal boundary, and an independently verified chain-complex identity. We then define an admissible relational-state model with that sequence boundary, effective contact edges, and stochastic histories. Second, we construct a gauge-invariant four-state observable and prove its invariance under local $\mathbb{Z}_2$ transformations. Third, we integrate annealed contacts exactly to obtain a sequence-dependent nonlocal effective interaction. Fourth, we examine an independently controlled smooth-field discretization, derive a simple uncoupled composition-field limit from the sequence law itself, and quantify the incompatibility of nonlinear readout and averaging. Fifth, we test the implemented finite-model identities with author-conducted Python computations, derive a fixed-support stability bound for contact-derived topological filtrations, and assess computational complexity, large-scale inference, and relationships to established genomic methods. We additionally distinguish microscopic-to-observable and data-to-complex maps in Section 32, including exact information-loss and prediction-value criteria. The fundamental GFT condensate and its continuum geometry remain a separate, conditional extension, not a numerical result of the present paper.

Figure 1 is a typed, branching navigation map: solid connections represent defined constructions or proved/verified relationships; dashed connections identify mechanisms still to be derived. The verified model begins with a designated two-complex support and a four-state terminal ensemble; the microscopic group-field mechanism that would generate this support is separately identified as

One specified terminal model; two unestablished physical links

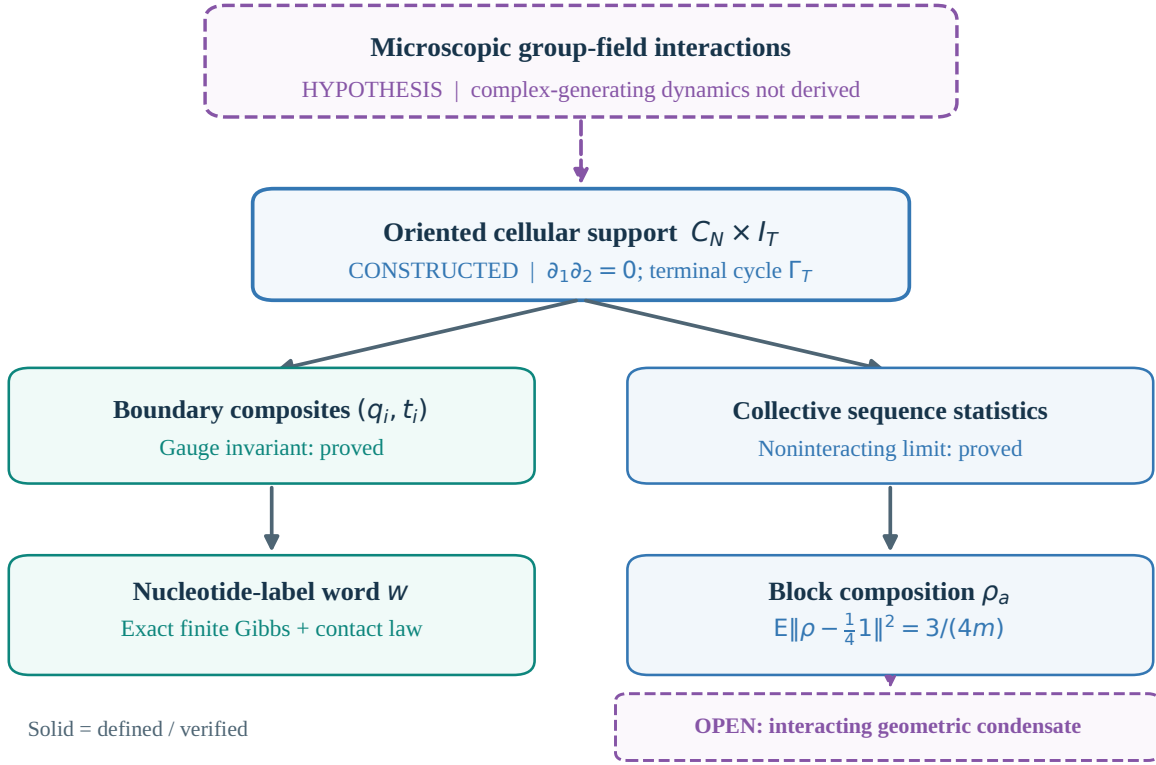


Figure 1: End-to-end model architecture with epistemic status attached to each stage. The oriented history two-complex is constructed exactly, terminal boundary invariants and their gauge properties are proved, and the normalized stochastic contact model is sampled. A microscopic group-field origin and a nontrivial interacting geometric condensate remain hypotheses. The boxes denote distinct spaces and maps; the figure does not assert that an ordinary four-letter string already carries a Lorentzian metric.

a research obligation. This separation prevents a diagram from silently becoming evidence for a speculative dynamical claim.

2 Representational Spaces and Observable Maps

2.1 Contour Order, Polymer Embedding, and Cellular History

A word w has a one-dimensional order; chromatin occupies ordinary physical space; and a history of changes may be represented by an independently defined combinatorial complex. Each notion of “spacetime carries its own observables and coordinates.” Figure 2 aligns the three descriptions while showing why each requires its own observables and coordinates. The physical cell state at position i may include w_i , an embedding $\mathbf{r}_i \in \mathbb{R}^3$, local biochemical attributes m_i , and nonlocal contacts c_{ij} . Only the symbolic and contact degrees of freedom occur in the exactly simulated Hamiltonian of Section 6.

There are also two epistemic directions. Forward generation asks for the distribution of observable boundary words given admissible histories and their physical weights. Inverse reconstruction asks which histories and internal degrees of freedom remain compatible with a measured boundary. An exact forward map can be many-to-one; gauge redundancy contributes one form of nonuniqueness and physically distinct latent contacts contribute another.

A Genomic contour i

Sequence index

**B Polymer embedding $\mathbf{r}(i, t) \in \mathbb{R}^3$**

Spatial proximity

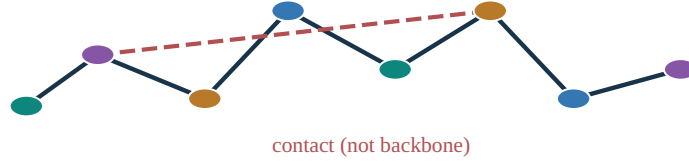
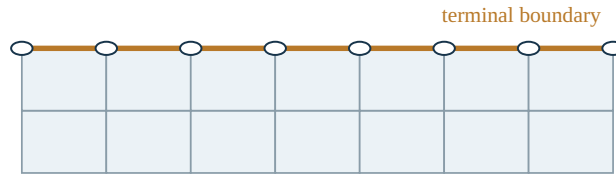
**C Cellular history index τ** Faces f_i^τ Contour i , biochemical time t , and history index τ are different variables.

Figure 2: Three distinct geometrical structures. Panel A: an ordered genomic contour carrying discrete symbols. Panel B: a schematic spatial polymer with an interaction between sequence-distant sites, projected onto the page. Panel C: a two-dimensional cellular support whose faces represent elementary compatibility or transition patches and whose terminal boundary bears the contour. None of these diagrams identifies the genomic contour coordinate with elapsed physical time.

3 Cellular Two-Complexes and the Genomic Boundary

3.1 A Product Cellulation of Finite Generative Histories

We begin with an explicit oriented CW two-complex that specifies the cellular meaning of “spacetime.” Let $N \geq 3$ and $T \geq 1$. Define the cellular support

$$\mathcal{K}_{N,T} = C_N \times I_T, \quad (1)$$

where C_N denotes an oriented cycle of N edges and I_T a path with T steps. The geometric realization is a cylinder, viewed in Fig. 3 after cutting and unrolling a fundamental domain. Its 0-cells v_i^τ have $i \in \mathbb{Z}/N\mathbb{Z}$ and $0 \leq \tau \leq T$; 1-cells are contour edges a_i^τ and transition edges b_i^τ ; 2-cells are squares f_i^τ , for $0 \leq \tau < T$. The coordinate τ here is an index for the discretized history; interpreting it as physical time requires a separately calibrated kinetic law.

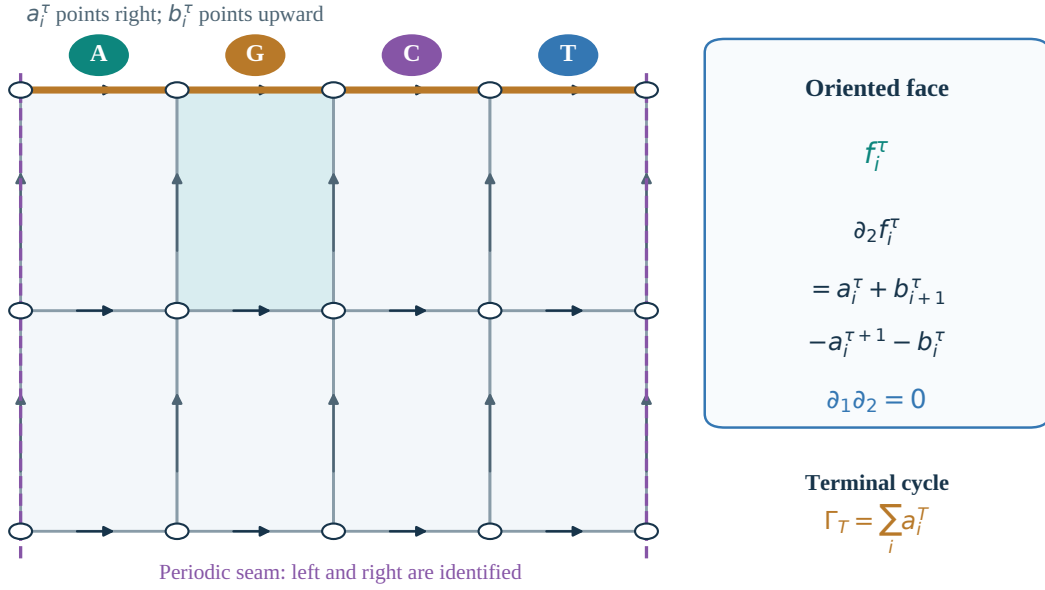
Definition 3.1 (Cellular incidence operators). *On the free integral cellular chain groups $C_k(\mathcal{K}_{N,T}; \mathbb{Z})$ define*

$$\partial_1 a_i^\tau = v_{i+1}^\tau - v_i^\tau, \quad \partial_1 b_i^\tau = v_i^{\tau+1} - v_i^\tau, \quad (2)$$

$$\partial_2 f_i^\tau = a_i^\tau + b_{i+1}^\tau - a_i^{\tau+1} - b_i^\tau. \quad (3)$$

Indices on the cycle are understood modulo N .

Proposition 3.2 (Cellular consistency and terminal boundary). *The cellular operators satisfy $\partial_1 \partial_2 = 0$. There are $N(T+1)$ vertices, $N(2T+1)$ edges, and NT faces. The top and bottom*

Oriented two-complex $C_4 \times I_2$ (unrolled cylinder)

Cellular support is prescribed; a field-theoretic derivation remains a separate task.

Figure 3: Explicit oriented two-complex $C_N \times I_T$ shown as an unrolled cylinder. Vertical sides are identified, producing periodic contour slices. Horizontal edges a_i^τ describe contour order and vertical edges b_i^τ connect successive history slices. Each square is a genuine two-cell, with an oriented cellular boundary. The highlighted upper cycle is the terminal genomic boundary; colored letters illustrate a readout assigned to its edges. The presence of this cellular structure is constructive, not evidence for an actual genomic spacetime condensate.

contours $\Gamma_T = \sum_i a_i^\tau$ and $\Gamma_0 = \sum_i a_i^0$ are closed cycles. With the face orientation of Eq. (3),

$$\partial_2 \sum_{\tau=0}^{T-1} \sum_{i=0}^{N-1} f_i^\tau = \Gamma_0 - \Gamma_T. \quad (4)$$

The Euler characteristic is zero and the Betti numbers are $(b_0, b_1, b_2) = (1, 1, 0)$.

Proof. Applying ∂_1 to any face in Eq. (3) cancels all four corner vertices in pairs. Summing all face boundaries cancels every vertical edge and all interior horizontal edges, leaving the initial and terminal cycles with opposite orientations. The cell counts follow directly from the product cellulation. The cylinder deformation retracts onto C_N , giving the stated Betti numbers by standard cellular-homology arguments (Hatcher, 2002); alternatively, ranks of the two incidence matrices give the same result. $\square$

This proposition establishes that the proposed visual cellular support is a mathematically valid two-complex. In the tested system, the topology and time slicing are specified as inputs before the boundary state is sampled; deriving this support from the contact-coupled Hamiltonian remains a separate research task.

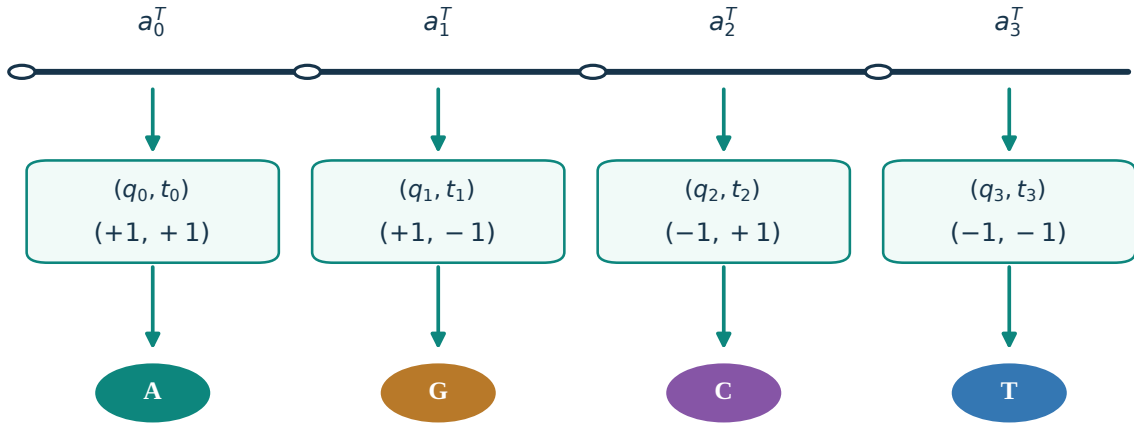
3.2 Sequence Attachment to the Terminal Boundary

The model's sequence lies on the oriented terminal contour Γ_T , with a specified edge-to-site correspondence. Choose the edge-to-site identification

$$\iota_T : a_i^T \mapsto i, \quad w_i = B(\mathcal{O}(a_i^T)), \quad w = w_0 \cdots w_{N-1}. \quad (5)$$

In the exact gauge model, $\mathcal{O}(a_i^T) = (q_i, t_i)$, with $q_i = h_i U_i h_{i+1}$. Boundary vertex matter variables h_i and edge links U_i define the invariant composite q_i , while the additional gauge-neutral t_i provides four states in total. Figure 4 traces these operations for four terminal edges, matching the worked example below. The chemical names A, C, G, and T are assigned through an explicit modeling convention; no nucleotide identities follow from cell topology alone. The distinction

Gauge-invariant terminal edge readout



Four named toy states; nucleotide chemistry is not inferred from this identification.

Figure 4: A worked representation map from oriented terminal boundary edges to four-state local invariants and the corresponding observable word. The boundary establishes order, the composite field supplies a gauge-invariant local state, and a deterministic readout supplies the symbolic label. The illustration distinguishes a four-state mathematical encoding from an actual derivation of nucleotide chemistry.

between edge occupancy, symbol label, and cellular face is fundamental: two different terminal words can live on the same $\mathcal{K}_{N,T}$; nonlocal contacts can change without altering the chain complex; a structural insertion would instead require replacing the chosen cellulation or specifying an admissible subdivision history.

3.3 An Explicit Four-Site Boundary Alignment

The mapping admits a direct algebraic check. Consider $\mathcal{K}_{4,2}$, with 12 vertices, 20 edges, and 8 square faces. The terminal cycle consists of $a_0^2, a_1^2, a_2^2, a_3^2$. Set all four terminal vertex variables to $h_i = +1$ and choose $U = (+1, +1, -1, -1)$ and $t = (+1, -1, +1, -1)$. The invariant terminal pairs are $q_i = h_i U_i h_{i+1} = U_i$ and hence

$$(q_i, t_i) = ((+, +), (+, -), (-, +), (-, -)), \quad w = \text{AGCT}. \quad (6)$$

Under the nonconstant gauge transformation $\eta = (-1, +1, -1, +1)$, the vertex signs become $h' = \eta$ and the edge links become $U' = (-1, -1, +1, +1)$, yet every $q'_i = q_i$. The terminal word is therefore invariant even though the microscopic gauge representatives differ. The interior faces continue to satisfy Eq. (3) independently of these terminal variable assignments.

This example also makes a limitation explicit. A geometrically valid two-complex can carry multiple words, and a specified word admits multiple gauge representatives. The cellular topology supplies the ordered *support*; chemical identity and generative history require additional constitutive and dynamical information. The finite model allows one to enumerate compatible representations; identifying which of them corresponds to a physical molecular process requires a calibrated state-transition law.

3.4 Boundary Laws and Algorithmic History Extensions

The finite Gibbs law in Section 6 is a distribution $\mu(x, c)$ over terminal states and contacts. Its extension to a chosen T -slice cellular support may be specified using a Markov transition kernel P on these states. If $\mu P = \mu$, then

$$\mathbb{P}(x^0, c^0, \dots, x^T, c^T) = \mu(x^0, c^0) \prod_{\tau=0}^{T-1} P((x^\tau, c^\tau), (x^{\tau+1}, c^{\tau+1})) \quad (7)$$

is normalized and has terminal marginal $\mu(x^T, c^T) = \mu(x, c)$. The proof is repeated summation using $\mu P = \mu$. A complete Gibbs sweep is such a kernel for the finite contact model. This provides a precise, algorithmic probability distribution over sequences of cellular slices. The construction specifies an algorithmic history kernel. A microscopic account of DNA synthesis, repair, or evolution would require physical kinetics connected to Eq. (22).

4 Relational Genomic Configurations and Generative Histories

4.1 Contour Coordinates, Spatial Contacts, and Generative Time

The word “spacetime” can easily become misleading in a genomic model. We distinguish: (i) *contour order*, an oriented polymer or genomic coordinate; (ii) *spatial contact*, an embedding-dependent physical relation among contour sites; (iii) *process history*, a partially ordered or parameterized collection of chemical and structural events; and (iv) *emergent effective geometry*, which requires a demonstrable scaling limit or reconstruction. Contour separation is not physical elapsed time. A Monte Carlo sweep is not necessarily physical time, and a relational edge is not automatically a Lorentzian causal interval.

Definition 4.1 (Finite relational genomic prototype). *Fix $N \geq 4$. Let $V_N = \mathbb{Z}/N\mathbb{Z}$ and let $E_{\text{bb}} = \{(i, i+1) : i \in V_N\}$ denote an oriented periodic backbone. Let $E_\star$ be a specified set of unordered non-nearest-neighbor pairs. A relational configuration is $\mathcal{K}(c) = (V_N, E_{\text{bb}}, E(c))$, where $E(c) = \{(i, j) \in E_\star : c_{ij} = 1\}$ and $c_{ij} \in \{0, 1\}$. The base orientation defines the order in which sequence symbols are read.*

The periodic choice is a mathematically convenient boundary condition; biologically linear chromosomes call for an interval contour. This tested terminal submodel is a graph (a one-dimensional complex with chords). Section 3 additionally constructs a valid two-complex that can support a history of such boundary states, but it is not yet produced by group-field dynamics or endowed with a physical causal metric. Linear backbones and strand-reversal conventions are possible extensions. The random contact edges are auxiliary effective interactions; we do not identify them directly with covalent DNA bonds or specific chromatin loops.

Definition 4.2 (Generative history). *A history $\mathcal{H} = (X_0, r_1, X_1, \dots, r_T, X_T)$ consists of admissible configurations X_t and applicable transformations $r_t : X_{t-1} \mapsto X_t$. When only state/contact updates are allowed, $\mathcal{H}$ is a path in a finite-state Markov system; when insertion, deletion, or replication are enabled, the object set and boundary cardinality can change. The family of all histories is generally a category of transitions rather than a single symmetry group.*

A physical process would require calibrated transition rates and, for active biological processes, possibly violations of detailed balance. In this draft we use equilibrium sampling as an analytically tractable reference. Its Monte Carlo history is algorithmic; measured replication and developmental trajectories remain distinct empirical objects.

4.2 A Normalized Joint Law for Structures and Sequences

For latent discrete complexes $\mathcal{K}$, internal variables ξ , environmental conditions e , and a boundary word w , a general target distribution is

$$\mathbb{P}_\theta(w, \mathcal{K}, \xi \mid e, N) = Z_\theta(e, N)^{-1} W_\theta(\mathcal{K}, \xi; e, N) \prod_{i \in \partial \mathcal{K}} q_\theta(w_i \mid \mathcal{O}_i(\mathcal{K}, \xi; e)). \quad (8)$$

We require $W_\theta \geq 0$, $Z_\theta < \infty$, and $\sum_{a \in \Sigma} q_\theta(a \mid \mathcal{O}) = 1$. For a deterministic readout, q is an indicator function. An end-to-end generative model samples $\mathcal{K}, \xi$ and then w . The inverse problem conditions on observed w , contacts, or other measurements and may be many-to-one.

5 Gauge-Invariant Composite Variables and Four-State Readout

5.1 Local Gauge Action and Gauge-Invariant Bond Variables

We begin with a deliberately small finite lattice gauge construction, in the spirit of the standard distinction between lattice gauge redundancies and observable spin degrees of freedom (Kogut, 1979). At each vertex place a charged variable $h_i \in \{\pm 1\}$, and on each oriented backbone edge a link variable $U_i \in \{\pm 1\}$. Local gauge transformations are $\eta_i \in \{\pm 1\}$ acting as

$$h_i \mapsto h'_i = \eta_i h_i, \quad U_i \mapsto U'_i = \eta_i U_i \eta_{i+1}. \quad (9)$$

Define the gauge-neutral bond observable

$$q_i := h_i U_i h_{i+1} \in \{\pm 1\}, \quad q'_i = q_i. \quad (10)$$

A second, gauge-neutral binary variable $t_i \in \{\pm 1\}$ gives a pair $x_i = (q_i, t_i)$, with four physical labels. Define the bijective encoding

$$(+, +) \mapsto A, \quad (+, -) \mapsto G, \quad (-, +) \mapsto C, \quad (-, -) \mapsto T. \quad (11)$$

This defines symbolic labels for four abstract states. Molecular nucleotide identities, hydrogen-bond energies, and chemical interactions require a separate molecular model. Under the local gauge action, x_i and hence $w_i = B(x_i)$ remain unchanged.

Proposition 5.1 (Exact local gauge invariance). *For any $\eta \in (\mathbb{Z}_2)^N$, the four-symbol sequence w obtained from Eqs. (10)–(11) is unchanged under Eq. (9). Any Hamiltonian depending only on (x, c) is gauge-invariant.*

Proof. The transformed bond product is $q'_i = (\eta_i h_i)(\eta_i U_i \eta_{i+1})(\eta_{i+1} h_{i+1}) = q_i$ because $\eta_i^2 = 1$. The variable t_i is invariant by construction, so the pair, its bijective readout, and every function of (x, c) are invariant. $\square$

On the circle, the Wilson product $\prod_i U_i = \prod_i q_i$ is gauge-invariant. Every assignment of q has exactly 2^N pairs of (h, U) representing it, since any choice of h fixes $U_i = h_i q_i h_{i+1}$. This constant gauge multiplicity cancels from normalized probabilities.

Proposition 5.2 (Exact gauge quotient and its explanatory limit). *For the periodic N -edge terminal system, the action of $(\mathbb{Z}_2)^N$ on (h, U, t) is free and the invariant map*

$$[(h, U, t)] \mapsto (q, t), \quad q_i = h_i U_i h_{i+1}, \quad (12)$$

is a bijection between gauge orbits and $\{\pm 1\}^{2N}$. Consequently, if the energy and observables depend only on (q, t, c) , introducing (h, U) leaves every normalized observable prediction of the four-state boundary model invariant.

Proof. Surjectivity follows by choosing $h_i = 1$ and $U_i = q_i$. Given two configurations with identical (q, t) , take $\eta_i = h'_i h_i$; the equality of bond products then forces $U'_i = \eta_i U_i \eta_{i+1}$. Hence they are gauge-related, proving injectivity on orbits. The action is free because $h'_i = h_i$ forces every $\eta_i = 1$. Each orbit has 2^N elements, and its multiplicity cancels between numerator and partition function for any invariant observable. $\square$

This exact equivalence defines the scope of the gauge construction: the present gauge lift is an exactly redundant parametrization of the terminal four-state model. A nontrivial gauge-mediated generative mechanism would require an independently justified gauge action, physical couplings, and gauge-invariant predictions not already built into the four-state boundary law.

Two gauge representatives; identical four-symbol boundary readout

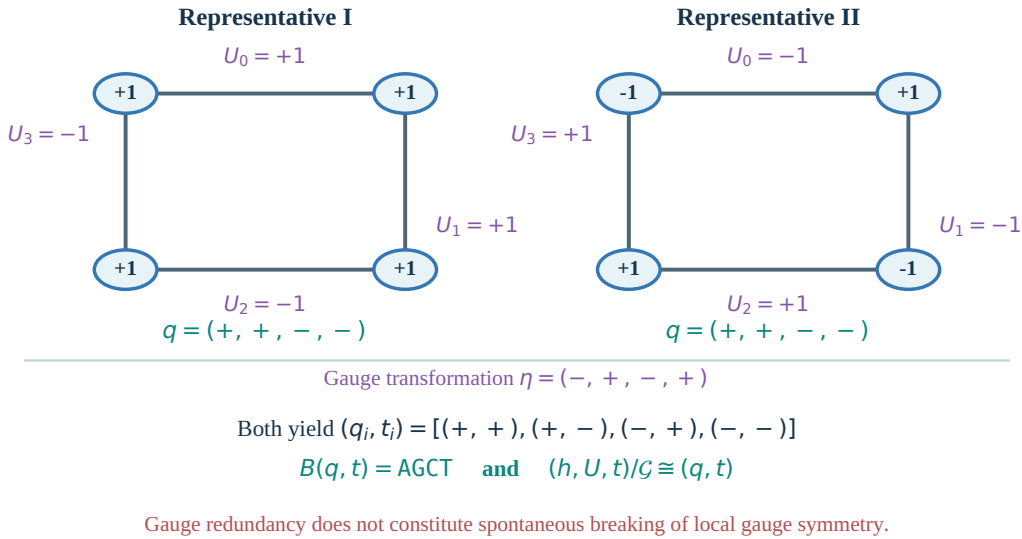


Figure 5: Two explicit four-vertex gauge representatives related by the transformation $\eta = (-1, +1, -1, +1)$ give the same bond-invariant sequence $q_i = h_i U_i h_{i+1}$. Gauge-dependent variables change while the physical readout is unchanged. Accordingly, symbolic state selection in this toy model is not a spontaneous breaking of local gauge redundancy.

5.2 Local Gauge Invariance and Physical State Selection

Elitzur's theorem excludes conventional spontaneous breaking of a local gauge redundancy in an unfixed gauge-invariant lattice theory (Elitzur, 1975). Therefore, the title phrase “via Gauge Symmetry Breaking” identifies the proposed theoretical mechanism being investigated, rather than a process established in this manuscript. The actual finite-model mechanism is gauge-invariant composite readout with a distinct global symmetry and possible *physical* state selection.

The implemented readout uses invariant composites; it neither breaks a local gauge symmetry nor establishes a Higgs phase. The equal-weight four-label theory below has a separate global permutation symmetry, which must be distinguished from Eq. (9).

A soft local four-component, gauge-invariant field $\Phi_i \in \mathbb{R}^4$ can supply an idealized state-selection potential,

$$V(\Phi) = -\mu_0^2 \|\Phi\|^2 + \lambda \|\Phi\|^4 + \kappa \sum_{a < b} \Phi_a^2 \Phi_b^2, \quad \mu_0^2, \lambda, \kappa > 0. \quad (13)$$

Its minimum set consists of eight points $\{\pm v e_a : a = 1, \dots, 4\}$, with $v^2 = \mu_0^2/(2\lambda)$; imposing the nonnegative orthant yields four minima. The elementary derivation is given in Appendix C. These minima supply a possible four-state phenomenological readout, *if* a nonnegative constraint or a different physical potential is specified. The numerical tests use the exactly discrete four-state reduction, not direct sampling of Eq. (13).

6 A Normalized Contact-Coupled Sequence Model

6.1 Contact-Coupled Hamiltonian and Partition Function

Let $\delta_{ij}(x) = \mathbf{1}\{x_i = x_j\}$. On the finite relational complex define

$$H_N(x, c) = -J \sum_{i \in V_N} \delta_{i, i+1}(x) + \sum_{(i, j) \in E_\star} c_{ij} [\nu - K \delta_{ij}(x)], \quad (14)$$

where $J \geq 0$ is a contour preference, $K \geq 0$ the contact-mediated matching reward, ν the contact activation cost, and $\beta > 0$ inverse effective temperature. The measure of the finite model is

$$\mathbb{P}_{N, \beta}(x, c) = Z_N^{-1} \exp[-\beta H_N(x, c)], \quad Z_N = \sum_{x \in \{0, 1, 2, 3\}^N} \sum_{c \in \{0, 1\}^{|E_\star|}} \exp[-\beta H_N(x, c)]. \quad (15)$$

By construction, this model includes a sequence output, an annealed contact graph, and well-defined generation and inference distributions. Its Potts-like label symmetry is an idealization and does not describe the unequal biochemical properties of A, C, G, and T.

Proposition 6.1 (Exact elimination of contact edges). *For any fixed x , contact variables are conditionally independent and*

$$\mathbb{P}(c_{ij} = 1 \mid x) = \frac{1}{1 + \exp\{\beta[\nu - K \delta_{ij}(x)]\}}. \quad (16)$$

The marginal sequence distribution is

$$\mathbb{P}(x) = Z_N^{-1} \exp\left(\beta J \sum_i \delta_{i, i+1}(x)\right) \prod_{(i, j) \in E_\star} \left[1 + e^{\beta(K \delta_{ij}(x) - \nu)}\right]. \quad (17)$$

Proof. For each edge, the binary sum is $\sum_{c=0}^1 e^{-\beta c(\nu - K \delta)} = 1 + e^{\beta(K \delta - \nu)}$. Distinct edge factors separate given x ; dividing the active term by their sum gives Eq. (16). $\square$

The induced effective Hamiltonian, up to an additive constant, is

$$H_{\text{eff}}(x) = -J \sum_i \delta_{i, i+1}(x) - \beta^{-1} \sum_{(i, j) \in E_\star} \log\left[1 + e^{\beta(K \delta_{ij}(x) - \nu)}\right]. \quad (18)$$

A contact edge gives an additional agreement reward

$$\Delta_{\text{contact}} = \beta^{-1} \log \left[\frac{1 + e^{\beta(K - \nu)}}{1 + e^{-\beta \nu}} \right] > 0 \quad (K > 0). \quad (19)$$

This is an explicit example of sequence-level nonlocal coupling induced by summing over latent relational degrees of freedom. The presence of such a coupling does not by itself prove that the resulting string language requires a context-sensitive grammar; formal-language classification would need its own theorem.

Annealed nonlocal contacts on a fixed genomic contour

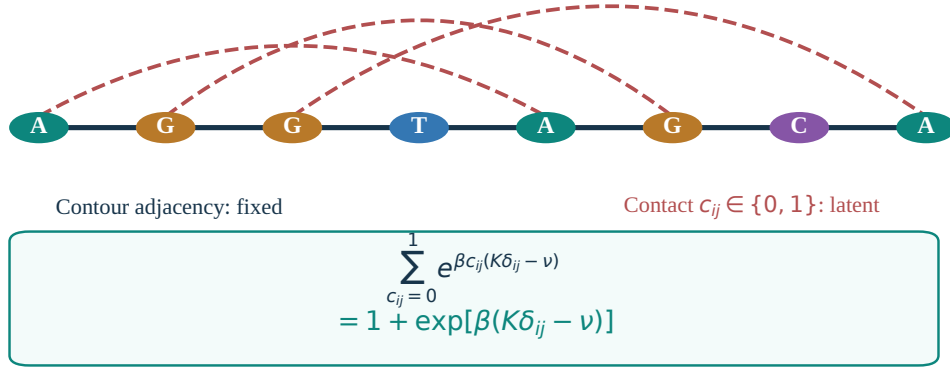


Figure 6: Physical mechanism for effective contextual dependence: terminal symbols occupy the genomic contour, while optional annealed contact edges favor interactions between nonadjacent labels. Summing over the latent binary contacts yields a product of sequence-dependent factors in Eq. (17). Dashed or arched contacts are auxiliary couplings, not additional faces of the cellular history complex.

6.2 Global Permutation Symmetry and Finite-Volume Constraints

Because Eq. (14) depends only on equality indicators, a single global permutation $\pi \in S_4$ acting as $x_i \mapsto \pi(x_i)$ preserves H_N and the measure. Hence

$$\mathbb{P}(x_i = a) = 1/4 \quad \text{for every } i \text{ and } a, \quad (20)$$

for the unconstrained finite model. Ordered finite configurations are possible, but the expectation of a signed symmetry-breaking order parameter vanishes in the symmetric ensemble. For a one-dimensional finite-state system whose interaction range and couplings remain uniformly bounded as $N \rightarrow \infty$, standard finite-range transfer-matrix arguments exclude an ordinary positive-temperature symmetry-breaking phase transition. The statement need not hold for arbitrary N -dependent long-range contact sets. Therefore increasing finite-size coherence in a numerical example should not be called proof of a thermodynamic condensate.

7 Joint Sampling of Sequences and Relational Contacts

The distribution in Eq. (15) admits a Gibbs sampler. Conditional on contacts and all other labels, the label at site i has four possible weights,

$$\mathbb{P}(x_i = a \mid x_{-i}, c) \propto \exp \left\{ \beta \left[J(\mathbf{1}\{a = x_{i-1}\} + \mathbf{1}\{a = x_{i+1}\}) + K \sum_{j:(i,j) \in E_\star} c_{ij} \mathbf{1}\{a = x_j\} \right] \right\}. \quad (21)$$

The activation cost does not depend on the label and therefore cancels from this conditional. After the site updates, each contact is resampled using Eq. (16). Each component update preserves the

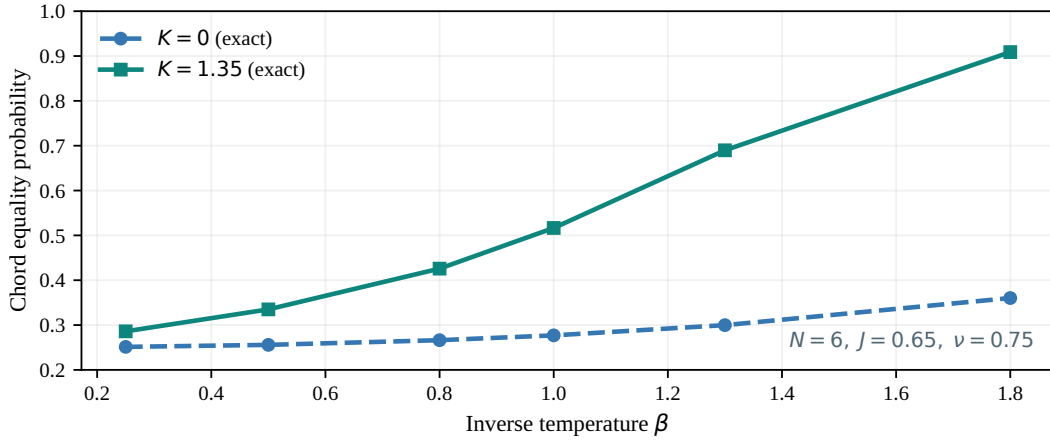


Figure 7: Exactly enumerated six-site model: equality probability for the specified chord pairs, with and without contact coupling, as inverse temperature varies. The curves are exact finite-model calculations, not genomic measurements.

target measure, and the full chain is irreducible for finite parameters because every local label and edge value has positive conditional probability. With systematic or randomized sequential scans, it has the stationary distribution in Eq. (15). A finite run still requires mixing checks before any parameter-dependent claim.

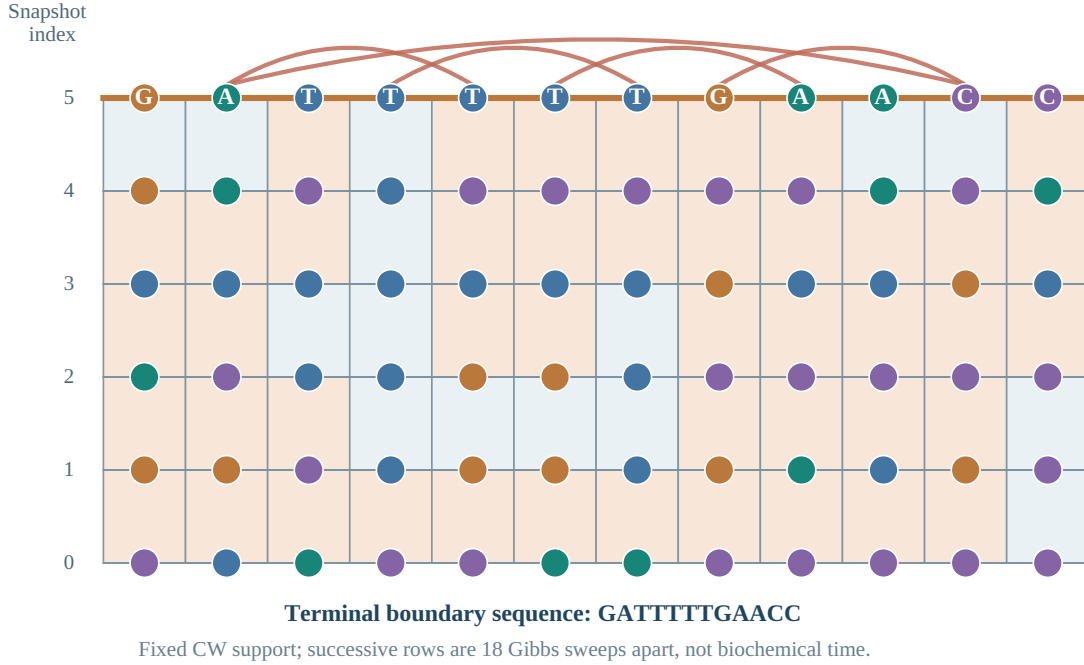


Figure 8: Monte Carlo configurations on the *fixed* cellular support $C_{12} \times I_5$. Each row consists of four-state variables after a prescribed number of Gibbs sweeps; shading marks individual cells whose associated local label changes between sampled rows. The highlighted terminal boundary carries the sampled 12-symbol word GATTTTGAACC, with sampled active contact arcs. This links the earlier schematic two-complex to an executable sequence/contact sampler; the vertical index denotes Gibbs-sweep sample index, not measured biochemical time or GFT interaction events.

An additional fully specified sampled cellular history is shown in Fig. 8. The stationary-approximate Gibbs update kernel decorates the fixed history complex and produces an observable terminal word together with contact arcs. It remains an algorithmic lifting of the finite terminal law rather than an interaction-generated cellular topology. The computational pipeline generates (x, c) , recovers the four-state observable $w = B(x)$, and calculates relational observables. For a model with explicit (h, U, t) , one may sample any gauge representative after drawing q ; the extra gauge degrees do not change w or gauge-invariant statistics. Conversely, an observed sequence fixes x in the toy model but not c , h , or U . This is a simple example of nonidentifiability of latent histories even when the observed alphabet is deterministic.

8 Group-Field Generativity and Condensate Conditions

8.1 Group-Field Interactions and Oriented Boundary Complexes

A possible generalization employs a field $\psi : G^d \times \mathcal{M} \rightarrow \mathbb{C}$ with group arguments representing gluing data and $\mathcal{M}$ representing chemical or structural degrees of freedom. A schematic action is

$$S[\psi, \bar{\psi}, \Phi] = \int \bar{\psi} \mathbf{K} \psi + \sum_{\alpha} \lambda_{\alpha} \mathbf{V}_{\alpha}[\psi, \bar{\psi}] + S_{\text{phys}}[\Phi] + S_{\text{couple}}[\psi, \Phi]. \quad (22)$$

Interaction kernels $\mathbf{V}_{\alpha}$ must be designed so that the perturbative combinatorial objects possess admissible oriented boundary paths and specified relational contacts. Gauge invariance, admissible topology, amplitudes and convergence cannot be inferred from notation alone. An abstract sum over generated complexes,

$$Z = \sum_{\mathcal{K} \in \mathfrak{K}} A_{\theta}(\mathcal{K}), \quad (23)$$

needs a defined class $\mathfrak{K}$, positive or quantum amplitudes as appropriate, and either convergence or controlled regularization. Our annealed contact model in Sections 6–7 is a tractable subfamily of weighted relational complexes, not a Feynman-diagram derivation from Eq. (22).

8.2 Cellular Amplitudes and the Missing Dynamical Derivation

To turn the cellular support of Section 3 into a predicted object, one must specify an interaction model that yields an actual measure on allowable complexes, not merely draw the complexes. For a nonnegative statistical model the missing object is a normalized cylinder amplitude $A_{\theta}(\mathcal{K}, x^0, x^T)$ or a family of elementary kernels satisfying an explicit composition rule. For a quantum amplitude model there are additional problems of interference, observable probabilities, and boundary conditions. Neither the notation $\sum_{\mathcal{K}} A_{\theta}(\mathcal{K})$ nor the occurrence of a cellular two-complex determines these structures. The distinction between different formulations and interpretations of group field theory remains an active foundational question (Mozota Frauca, 2024). In particular, the $\mathbb{Z}_2$ boundary gauge model introduced here should not be silently identified with the local Lorentz or rotation symmetries of gravity-inspired GFT.

A proposed route is to fix an admissible class of oriented cellular bordisms with genomic contour boundaries, attach gauge-invariant boundary observables and an action to each history, and test whether a projectively consistent family of finite cutoff measures can be normalized. Only after this step can a geometric large-system regime be investigated. The present article provides explicit boundary and cellular consistency tests, which are necessary construction checks but are insufficient for a full field-theoretic derivation.

8.3 Finite-Size Coherence, Thermodynamic Order, and Group-Field Condensation

A *finite collective state* may be summarized by correlations or the invariant maximum occupancy

$$Q_N(x) = \frac{4 \max_{a \in \Sigma} \rho_a(x) - 1}{3}, \quad \rho_a(x) = \frac{1}{N} \sum_{i=1}^N \mathbf{1}\{x_i = a\}. \quad (24)$$

The quantity $Q_N \in [0, 1]$ is a finite-size coherence statistic, not proof of a broken-symmetry phase. A *thermodynamic condensate or ordered phase* requires a specified limit, order of limits for any symmetry-breaking source, and appropriate long-range order or gauge-invariant observables. A *GFT condensate* is a particular many-body state of field quanta for which collective equations can sometimes be derived (Gielen et al., 2013, 2014). It is neither equivalent to Eq. (24) nor guaranteed by smoothing a discrete field.

A genomic interpretation would require evidence that selected collective variables reproduce polymer shape, contact statistics, replication or transcription behavior, and their environmental dependencies. The present finite model verifies none of those biological claims.

9 Smooth-Field Continuum Limit and Its Scaling Test

9.1 Smooth-Field Discretization and Continuum Energy

Let $\Phi : [0, 1] \rightarrow \mathbb{R}^m$ be a smooth periodic field, $a = 1/N$, and $\Phi_i = \Phi(ia)$. Define the discrete energy

$$\mathcal{F}_N(\Phi) = \frac{\kappa}{2a} \sum_{i=0}^{N-1} \|\Phi_{i+1} - \Phi_i\|^2 + a \sum_{i=0}^{N-1} V(\Phi_i), \quad \Phi_N = \Phi_0. \quad (25)$$

For smooth Φ and continuous V , the corresponding continuum functional is

$$\mathcal{F}(\Phi) = \int_0^1 \left[\frac{\kappa}{2} \|\partial_x \Phi\|^2 + V(\Phi) \right] dx. \quad (26)$$

Standard difference-quotient and Riemann-sum arguments yield $\mathcal{F}_N(\Phi) \rightarrow \mathcal{F}(\Phi)$ for a fixed smooth sampled field. A variational Γ -convergence claim for arbitrary variable configurations would require equicoercivity, boundary conditions, and separate proof; it is not claimed here.

Proposition 9.1 (Explicit second-order benchmark). *Let $\Phi(x) = (u + A \cos 2\pi x, v + A \sin 2\pi x)$, and $V(\Phi) = \frac{m^2}{2} \|\Phi\|^2$. For $N \geq 3$, the potential Riemann sum is exact and*

$$\mathcal{F}_N = 2\kappa A^2 N^2 \sin^2(\pi/N) + \frac{m^2}{2} (u^2 + v^2 + A^2), \quad (27)$$

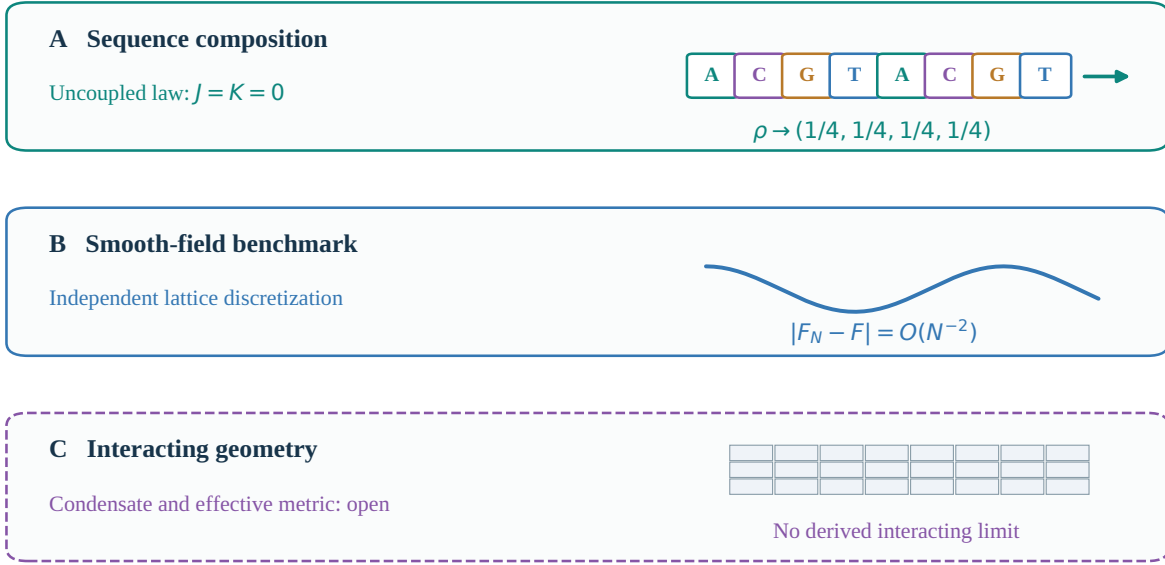
while

$$\mathcal{F} = 2\kappa A^2 \pi^2 + \frac{m^2}{2} (u^2 + v^2 + A^2). \quad (28)$$

Thus $|\mathcal{F}_N - \mathcal{F}| = \frac{2\kappa A^2 \pi^4}{3N^2} + O(N^{-4})$.

Proof. The squared adjacent difference of the periodic sinusoidal vector equals $4A^2 \sin^2(\pi/N)$, independent of site. The means of both sine and cosine vanish, and their squared means total one. Applying $\sin z/z = 1 - z^2/6 + O(z^4)$ gives the expansion. $\square$

Three continuum questions; different evidential status



Smoothness or a composition LLN is not evidence of a physical spacetime condensate.

Figure 9: Two mathematically different routes. Left: the four-state model has a provable composition law of large numbers in the uncoupled regime, while a separately specified smooth field admits a controlled lattice discretization. Right: a genuine interacting geometric condensate remains an unproved possibility requiring a collective-state construction, its dynamics, and gauge-invariant geometry observables. A smooth curve is not evidence of a physical condensate.

9.2 A Composition-Field Limit Derived from the Four-State Model

The smooth-field functional in Proposition 9.1 is separate from the stochastic contact model. A limited but direct continuum statement is nonetheless available inside the latter. Set $J = K = 0$, fixing any finite β and ν . Then the four-symbol variables x_i are independent and identically distributed with $\mathbb{P}(x_i = a) = 1/4$; the independent contact variables have no influence on their law. For a block of m symbols let $\rho_a^{(m)} = m^{-1} \sum_{i=1}^m \mathbf{1}\{x_i = a\}$.

Proposition 9.2 (Uncoupled composition-field law of large numbers). *In the noninteracting subfamily $J = K = 0$, the block composition vector $\rho^{(m)}$ obeys*

$$\mathbb{E}[\rho_a^{(m)}] = \frac{1}{4}, \quad \mathbb{E}\left[\sum_{a=1}^4 (\rho_a^{(m)} - \frac{1}{4})^2\right] = \frac{3}{4m}. \quad (29)$$

Consequently $\rho^{(m)} \rightarrow (1/4, 1/4, 1/4, 1/4)$ in mean square and in probability as $m \rightarrow \infty$.

Proof. The count vector has a multinomial distribution with m trials and four equally likely outcomes. Each component has variance $(1/4)(3/4)/m = 3/(16m)$ and the sum of component squared deviations has expectation four times that variance. Markov's inequality gives convergence in probability. $\square$

This is a law-of-large-numbers limit for block observables of the *same* symbolic model in its explicitly uncoupled regime, yielding only a trivial homogeneous composition field. It is not, by itself, a continuum spacetime limit. It neither establishes a nontrivial interacting continuum fixed point nor identifies genomic physical spacetime.

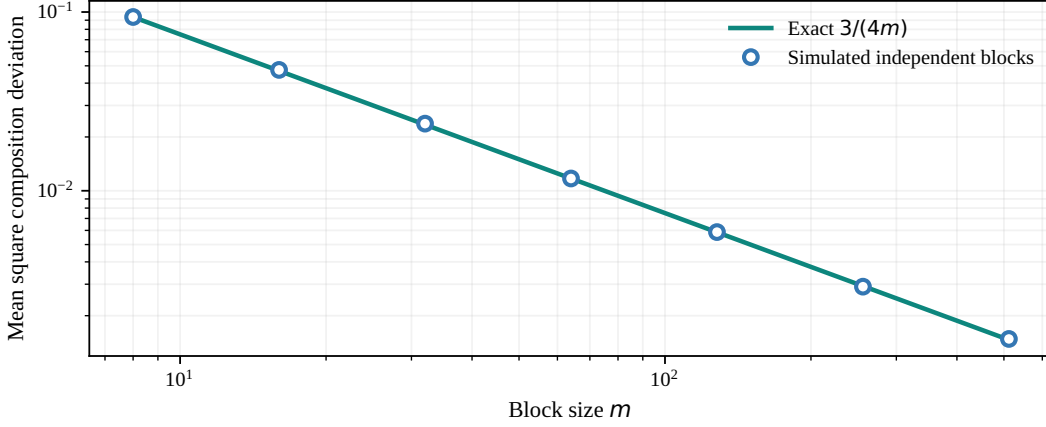


Figure 10: Composition-field benchmark derived from the noninteracting subfamily of the sequence model. Independent Monte Carlo blocks approach the exact $3/(4m)$ mean-square fluctuation law. The limiting field is spatially constant and does not establish an emergent metric or condensate.

The prototype lattice sequence itself is discontinuous and categorical. There is no canonical differentiable four-letter trajectory. A continuum description can instead concern composition fields, smoothed gauge-invariant observables, polymer embeddings, or other collective variables. Deriving their equations directly from a microscopic genomic stochastic process is a further task.

10 Noncommutativity of Readout and Coarse-Graining

Let $\Phi_i \in \mathbb{R}_+^4$ be a nonzero soft local state and define the deterministic readout $B(\Phi_i) = e_{\arg \max_a \Phi_i^a} \in \{e_1, e_2, e_3, e_4\}$, resolving ties by a fixed rule. For a block D of b sites, define the coarse-grained symbol histogram and field mean,

$$C_D^\Sigma(B(\Phi)) = \frac{1}{b} \sum_{i \in D} B(\Phi_i), \quad C_D^\Phi(\Phi) = \frac{1}{b} \sum_{i \in D} \Phi_i. \quad (30)$$

The two paths $C_D^\Sigma \circ B$ and $B \circ C_D^\Phi$ have a common codomain in the probability simplex when a one-hot vector is regarded as a degenerate distribution.

Proposition 10.1 (A readout/coarse-graining obstruction). *If a block contains at least two distinct winning labels, then $C_D^\Sigma(B(\Phi)) \neq B(C_D^\Phi(\Phi))$. If every local winner is the same unique label, then the two operations agree.*

Proof. With at least two winners, the left-hand side is a nontrivial convex combination of distinct vertices of the three-simplex, while the right-hand side is one vertex. If all unique local winners are the same a , then $\Phi_i^a > \Phi_i^b$ for every $b \neq a$ and every site, so their average has the same unique winner a ; both maps yield e_a . $\square$

For any block with winning-label frequencies $p_a = b^{-1} \sum_{i \in D} \mathbf{1}\{B(\Phi_i) = e_a\}$ and coarse-field winner $k = \arg \max_a b^{-1} \sum_{i \in D} \Phi_i^a$, the discrepancy obeys the exact identity

$$\|C_D^\Sigma(B(\Phi)) - B(C_D^\Phi(\Phi))\|_1 = 2(1 - p_k). \quad (31)$$

In particular the two paths agree exactly if and only if every site reads out as k ; heterogeneous local readouts cannot equal a single hard coarse-field label. For the explicit block $\Phi_i = e_i$ for $i = 1, \dots, 4$, the first path yields $(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4})$ and the second yields e_1 under the chosen tie rule. Their L^1 distance is 1.5. This failure is not a pathology: hard state selection discards uncertainty

and sub-block heterogeneity. A probabilistic effective readout $\bar{B}_\ell$ could improve correspondence, but it would need to be derived and calibrated, not postulated as automatically identical to the microscopic readout.

Noncommuting readout and coarse-graining maps

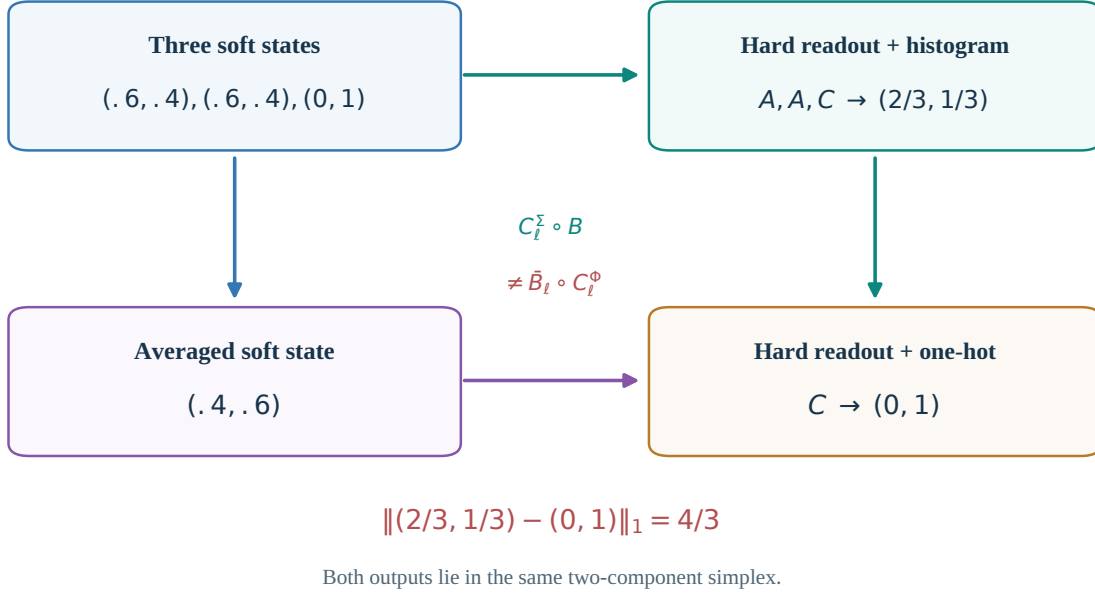


Figure 11: A tie-free counterexample to commutation. For three local soft states (0.6, 0.4), (0.6, 0.4), and (0, 1), local hard readout yields (A, A, C) and therefore the histogram $(2/3, 1/3)$, while averaging first yields (0.4, 0.6) and hard readout C . The resulting L^1 difference is $4/3$. This illustrates Proposition 10.1 without invoking unspecified coarse-graining operators.

11 Numerical Verification of the Finite Model

11.1 Integer Incidence Checks for the Oriented Two-Complex

We implemented the oriented product cellulation using integer-valued boundary matrices, independently of the Potts-like boundary model. For $N \in \{3, 6, 8, 12\}$ and $T \in \{1, 2, 3, 4\}$ (paired in order), direct matrix multiplication gives $\partial_1 \partial_2 = 0$ exactly and verifies both terminal cycles. The same calculations recover Euler characteristic zero and Betti numbers (1, 1, 0); Table 6 (Appendix A) reports the test cases. Since all incidence coefficients are integers, these are exact finite computations rather than approximate floating-point convergence claims.

The separate boundary-readout test reconstructs a randomly sampled four-letter terminal word from a gauge-equivalent bond representation without changing any letter. It confirms compatibility between the chosen terminal edge indexing and the gauge-invariant mapping, while the dynamical generation of the two-complex remains a separate problem.

11.2 Exact Enumeration, Gibbs Sampling, and Verification Protocol

All experiments were performed in Python with NumPy and Numba; Matplotlib was used for figures. The following author-conducted checks were run using exploratory Python scripts. A corresponding public release of this paper-specific implementation is not asserted; the methods, parameter values, and results are documented here, and implementation requests may be directed

to the author as explained in the code-availability statement. Exact enumeration uses $4^6 = 4096$ sequences on a six-site periodic backbone with candidate chord edges $E_\star = \{\{i, i+2\} : i \in \mathbb{Z}/6\mathbb{Z}\}$ interpreted as a set (six distinct edges). Model parameters are $\beta = 1$, $J = 0.65$, $K = 1.35$, $\nu = 0.75$. We directly sum over all 2^6 contact patterns for 800 randomly selected sequences to verify Proposition 6.1 against the analytic product formula. Gauge invariance is checked under 3,000 independently sampled transformations with random chain lengths between 4 and 69.

For Monte Carlo, we use 130,000 full sweeps, discard 10,000, and retain every 12th sweep, giving 10,000 retained measurements. We calculate standard errors from 25 contiguous batch means; these are estimates rather than a substitute for a formal mixing bound. A second, illustrative finite-size experiment uses $N = 40$, chord offset four, 16,000 sweeps at each inverse temperature, 4,000 discarded, and every 16th retained. We do not use these runs to estimate a thermodynamic critical point.

11.3 Exact Contact Marginalization and Monte Carlo Agreement

Random gauge transformations produced zero failures among 3,000 tests. Direct binary contact summation and Eq. (17) agreed with maximum observed relative error 1.90×10^{-15} , consistent with double-precision roundoff. The exact-versus-Monte-Carlo observable comparison, including batch standard errors, was retained in the author’s numerical records; the key quantitative checks are summarized here.

The largest discrepancy is 1.59 batch standard errors. Such agreement is consistent with a correctly implemented Gibbs sampler and weight calculation, not a demonstration of experimentally realistic parameters. Exact chord agreement increases when $K > 0$, as shown in Fig. 7. At $\beta = 1$ the exact values are 0.276980 without the contact coupling and 0.516525 with $K = 1.35$.

11.4 Second-Order Continuum Convergence and Readout Noncommutativity

For Proposition 9.1 we take $u = 0.55$, $v = 0.45$, $A = 0.2$, $\kappa = 1.3$, $m^2 = 0.8$. The closed-form continuum reference is $\mathcal{F} = 1.2444388577$. The full size sweep and absolute-error values were retained in the author’s numerical records; Fig. 25 in Appendix B displays the independently constructed benchmark.

A log-log fit across 12 lattice sizes gives an empirical slope of -1.998279 , agreeing with the -2 asymptotic rate in Proposition 9.1. The independently derived four-state composition check used 5,000 independent blocks at each of seven block sizes; the largest relative discrepancy from the exact $3/(4m)$ expected mean-square fluctuation was 1.295%, consistent with sampling fluctuations. The smooth-field discretization and the independent block-composition check assess different statements. Neither establishes a renormalization-group fixed point or convergence of the interacting categorical four-state model to the independently specified smooth field. Proposition 9.2 addresses only the exactly noninteracting composition limit, verified independently against its exact variance formula (Fig. 10).

For Proposition 10.1, we sampled 1,600 blocks containing four independent soft states with Dirichlet-distributed four-component vectors (shape parameters 0.45). The two readout paths differed in 98.3125% of blocks; mean L^1 mismatch was 1.012813. The 27 blocks having a single local winning label all satisfied equality, as predicted. These random tests illustrate the deterministic proposition and are not estimates of genomic heterogeneity.

Supplementary finite-size diagnostic. The one-seed $N = 40$ coherence run is treated as an exploratory diagnostic only. It does not support a phase-transition or condensate inference; no single-run coherence chart is retained in the main manuscript.

12 Computational Complexity of Finite Generative Inference

12.1 Exact Enumeration and Contact Marginalization

The existence of an exactly normalized finite distribution is logically separate from an efficient algorithm for evaluating its partition function or fitting it to observations. Let N denote the number of terminal four-state variables, $M = |E_\star|$ the number of candidate nonlocal contacts, and T the number of intervals in a fixed cylindrical cellular history. In the present finite model, direct enumeration of the joint configuration space $\{0, 1, 2, 3\}^N \times \{0, 1\}^M$ contains

$$|\Omega_{N,M}| = 4^N 2^M \quad (32)$$

configurations. Evaluating the unoptimized Hamiltonian for each takes $O(N + M)$ elementary contributions, yielding an $O((N + M)4^N 2^M)$ direct-enumeration bound. Equation (17) eliminates all binary contacts analytically: the remaining exact sum uses 4^N assignments and $O(N + M)$ work per assignment, or $O((N + M)4^N)$ time and $O(N + M)$ streaming workspace, excluding storage of the full distribution. These bounds describe a particular direct algorithm, not a universal lower bound on partition-function computation.

In fact, the model admits a stronger simplification than an arbitrary contact-dependent factor graph. Because $\delta_{ij} \in \{0, 1\}$, each marginal contact factor has only two values. Define

$$J_c(\beta, K, \nu) = \frac{1}{\beta} \log \frac{1 + \exp[\beta(K - \nu)]}{1 + \exp(-\beta\nu)}. \quad (33)$$

Then the exact sequence marginal can be written as

$$\mathbb{P}(x) = \frac{1}{\tilde{Z}_N} \exp \left\{ \beta J \sum_i \delta_{i,i+1}(x) + \beta J_c \sum_{(i,j) \in E_\star} \delta_{ij}(x) \right\}. \quad (34)$$

This is an ordinary four-state ferromagnetic Potts-type distribution on a contour-plus-contact graph when $J, K \geq 0$, with a coupling depending on the original contact parameters. All finite-model observables depending only on x can be computed from Eq. (34). In particular, neither the local gauge parametrization nor the auxiliary annealed binary contacts intrinsically reduces the asymptotic cost of generic exact inference. Their value at this stage is the explicit representation of observables and relations, not a demonstrated complexity advantage over the reduced model.

Proposition 12.1 (Exact reduction to a pairwise Potts model). *For fixed N , $E_\star$, $\beta > 0$ and parameters in Eq. (14), the sequence marginal in Eq. (17) coincides with Eq. (34). The multiplicative factor removed from its unnormalized weight is $(1 + e^{-\beta\nu})^M$, independent of x . For $K > 0$, $J_c > 0$; for $K = 0$, $J_c = 0$.*

Proof. For $\delta \in \{0, 1\}$, write $1 + e^{\beta(K\delta - \nu)} = (1 + e^{-\beta\nu}) \exp(\beta J_c \delta)$, where J_c is Eq. (33). Multiplying over M contacts gives the constant and the exponent in Eq. (34). Monotonicity of the exponential establishes the final claims. $\square$

12.2 Conditional Sampling, Graph Structure, and Mixing

For a fixed sparse candidate graph, a naive heat-bath sweep that inspects all M candidate contacts at each of the N terminal sites can have $O(NM + M)$ work. This describes a computational implementation choice rather than an inherent property of the probabilistic law. With indexed adjacency lists, a proposed state update at site i visits only its incident contour and candidate-contact neighbors. At fixed alphabet size $q = 4$, a full sequential sweep then costs $O(q(N + M))$, with a further $O(M)$ pass if the binary contact variables are explicitly resampled. Direct simulation of the marginalized law in Eq. (34) avoids contact sampling entirely when only sequence observables

are needed. Its static storage requirement is $O(N + M)$ with sparse edge lists, apart from outputs and intermediate diagnostics.

These are *per-sweep arithmetic and storage* bounds. They do not guarantee that a fixed number of sweeps produces an accurate independent sample. Near metastability or low-temperature ordering, integrated autocorrelation times can grow quickly, and local updates can mix poorly. The effective cost of estimating $\mathbb{E}[f]$ to specified uncertainty depends on burn-in, autocorrelation, variance, and target error. A rough diagnostic is $N_{\text{eff}} \simeq n_{\text{retained}}/(2\tau_{\text{int}})$ under stationary sampling and a conventional definition of integrated autocorrelation time; reliable error bars require checking those assumptions. The small-system batch comparisons in Section 11 do not provide genome-scale mixing guarantees. Cluster updates, tempering, particle methods, or other specialized samplers may be investigated, but only with appropriate conditions and validation.

Graph structure can make exact marginal inference feasible for restricted instances. If the pairwise interaction graph has treewidth w and an elimination order of induced width w is available, junction-tree or variable-elimination methods require time on the order of $O((N + M)q^{w+1})$ and messages exponential in w , up to representation-dependent polynomial factors. Treewidth is a property of the *full* interaction graph, not merely of the linear genomic backbone. Sparse long-distance contacts may still produce large treewidth; linear numbers of edges alone do not guarantee tractability. Repeated genomic motifs, local symmetry or shared parameters may reduce actual work in particular models, but no universal $|G|$ -fold parsing or partition-function speedup follows from group equivariance.

12.3 Cellular Representation, Topological Operations, and Memory

For the constructed product cellulation $C_N \times I_T$ in Section 3, the chain dimensions are

$$\dim C_0 = N(T + 1), \quad \dim C_1 = N(2T + 1), \quad \dim C_2 = NT. \quad (35)$$

An incidence-matrix implementation using dense arrays unnecessarily allocates products of these dimensions and can require $O(N^2T^2)$ storage. Every oriented 1-cell has two endpoint incidences and every square 2-cell has four boundary incidences, so sparse incidence storage requires only $O(NT)$ nonzero entries. Boundary products can be verified or applied without ever materializing dense matrices. For dynamic or topology-changing cellular histories, additions and deletions require careful maintenance of edge/face incidence and orientation; the fixed-cylinder formula is not a complexity theorem for such general histories.

Moreover, the candidate-contact set E_* in the current model is chosen externally and is not generated by the cellular faces. Treating all pairs of N sites as possible contacts would require $M = O(N^2)$ candidates. Sparse indexing improves only a genuinely sparse or truncated contact model; it cannot make a dense all-pairs physical interaction disappear. Generating contact proposals from a spatial index, distance cutoff, experimentally supported interaction list, or other justified sparsification rule is therefore a scientific modeling choice with an approximation error, not merely a programming optimization.

13 Large-Scale Genomic Analysis with Sparse and Multiscale Models

13.1 Distinct Meanings of Large-Scale Analysis

The phrase *genome-scale* covers at least four computationally different objectives: (i) reading and indexing a chromosome-scale nucleotide sequence; (ii) estimating an observable on a chromosome at coarser spatial or functional resolution; (iii) fitting latent relational and dynamical variables conditional on genomic observations; and (iv) reconstructing distributions over alternative structural or evolutionary histories. A method that achieves the first does not automatically accomplish

the remaining three. For multiple genomes, one must additionally distinguish shared pan-genomic coordinate structures from sample-specific variants, environments, and histories.

For perspective, an illustrative 3.1×10^9 -base genome would contain about 3.1×10^5 bins at 10-kb resolution. This reduces the number of *structural units* but does not make a four-letter readout at every original nucleotide available from each bin alone. A coarse bin should retain summary observables, uncertainties, and access to its high-resolution constituents as necessary. There are approximately 4.8×10^{10} unordered pairs among 3.1×10^5 bins; even one compact record per pair would make unfiltered all-pairs contacts burdensome. Neither the 10-kb choice nor the illustrative genome length is a validated operating point for our software. They demonstrate why a multiresolution graph is necessary for a prospective whole-genome realization.

13.2 Sparse Conditional Factors and Elimination of Redundant Variables

The immediate feasible computational core is the *reduced physical observable model*, rather than explicit sampling of all gauge representatives. The quotient result, Proposition 5.2, implies that enumerating gauge-related configurations contributes no new sequence information. The exact contact elimination in Proposition 12.1 similarly permits a collapsed sampler or factor-graph representation for the current equilibrium subfamily. This reduction should be performed before introducing sophisticated hardware or automatic differentiation.

A proposed implementation would store contiguous sequence blocks, candidate contact lists, condition-specific attributes, and sparse cellular incidence separately. On a static contact graph, local energy changes are calculated from incident neighbors. Larger chromatin structures could be represented by block-level state variables whose parameters are inferred from data rather than chosen by visual resemblance. If a dynamic polymer embedding $\mathbf{r}$ is added, contacts could be generated conditionally through $P(c \mid \mathbf{r}, e)$, and the spatial-neighbor index would need updating as the polymer moves. This is a different, more costly model than the exactly marginalizable contact factors of Section 6. Long-range electrostatic or topological effects may remain even after geometric truncation.

13.3 Effective Models Across Genomic Scales

Let $\pi_\ell : X \rightarrow Y_\ell$ be an explicit coarse-graining map from microscopic states X to scale- ℓ variables Y_ℓ . If a normalized microscopic law P_θ is available, its exact effective distribution is the pushforward

$$P_{\theta,\ell}(y) = \sum_{x:\pi_\ell(x)=y} P_\theta(x), \quad (36)$$

with the sum replaced by a measure integral when continuous degrees of freedom occur. An effective model $Q_{\varphi_\ell}(y)$ may approximate this marginal through fitted local factors, selected nonlocal interactions, or a reduced dynamical state. Equation (36) is a definition, not an efficient algorithm in itself. A computational approximation must control which observables are preserved and how error accumulates when repeated across scales.

For a set of observable tests $\mathcal{F}_\ell$, define a diagnostic discrepancy

$$\varepsilon_\ell(\mathcal{F}_\ell) = \sup_{f \in \mathcal{F}_\ell} \left| \mathbb{E}_{\pi_{\ell\#} P_\theta}[f] - \mathbb{E}_{Q_{\varphi_\ell}}[f] \right|. \quad (37)$$

Here $\pi_{\ell\#} P_\theta$ denotes the pushforward measure. An implementation should report discrepancies for concrete test classes, for example contact probabilities, motif statistics, selected boundary observables, and conditional perturbation responses. Consistency of two trained resolutions can then be tested without claiming that their inferred “metric” or physical units are already those of general relativity. A continuum limit, if any, additionally needs a scaling regime, convergence of a specified observable family, and potentially memory or nonlocal terms in the effective dynamics.

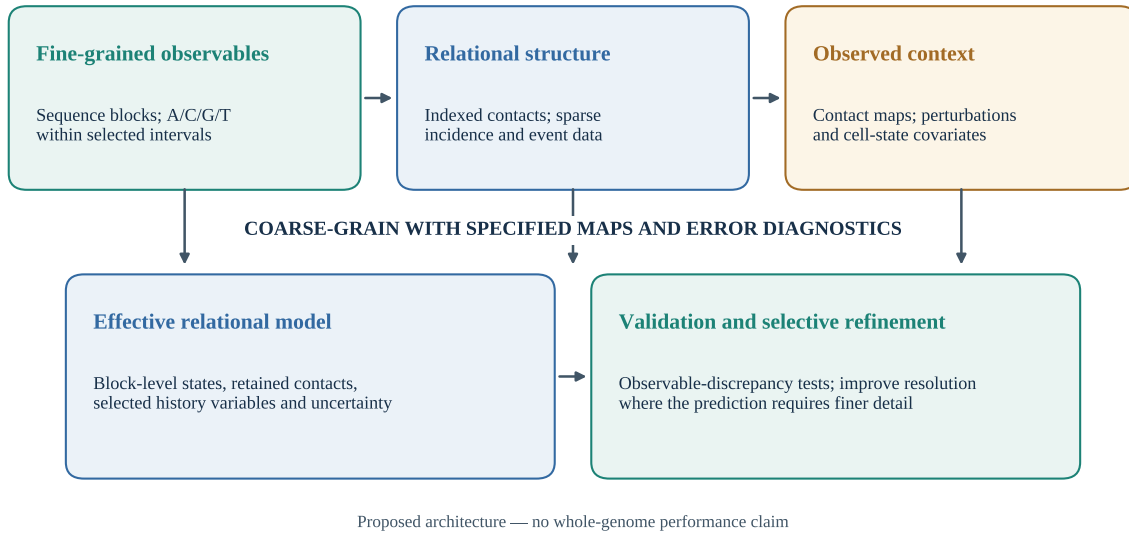


Figure 12: Prospective genome-scale computational organization, not a completed implementation. Fine sequence blocks, sparse contact candidates, and selected cellular histories are represented separately; coarse effective states are trained and tested against microscopic observables. Only regions requiring detailed predictions are refined. Each reduction introduces a measurable approximation error; the figure does not imply that computational cost becomes linear for arbitrarily nonlocal dynamics.

13.4 Current Feasibility and Research-Stage Extensions

Three feasibility tiers follow from these distinctions. *Currently verified:* exact small-system enumeration, finite-state sampling, fixed CW-complex chain checks, and independent toy continuum diagnostics. *Algorithmically feasible in principle but not benchmarked here:* larger fixed sparse interaction graphs with $O(N + M)$ local-sweep work, sparse cellulations, streaming sequence input, and hierarchical block representations. *Research-stage:* learning realistic genomic contact distributions, inferring topology-changing histories at scale, coupling sequence states to calibrated polymer dynamics, finding a controlled interacting continuum limit, and establishing a genuine condensate regime from a microscopic action. In particular, the present tests do not establish throughput, accuracy, memory use, or scientific validity on real chromosomes.

A concrete staged route would first replace full-graph scans with adjacency indexing; next introduce a fixed observed contact graph and conditional likelihood; then perform sparse inference on experimentally measured genomic windows; subsequently develop adaptive multiscale representations; and only then investigate richer dynamic complexes, polymer embeddings, and history inference. At each transition, simple methods must remain baselines, and the improved formalism must demonstrate an additional result rather than merely a more elaborate encoding.

13.5 Multi-Genome Coordinates and Shared Inference

For samples $s = 1, \dots, S$, a shared pangenome reference or sequence graph could store common intervals once and encode each sample through an annotated path, while sample-dependent physical conditions and inferred histories remain separate conditional variables. A shared index may reduce redundant parsing and storage when many samples reuse the same genomic structures, but it does not collapse unrelated environmental states or evolutionary histories into one universal 2-complex. In particular, a recombination graph describes ancestral relationships, whereas the present cellular history describes specified generative operations and boundary states; connecting them requires an explicit event correspondence rather than an identification of graph vertices.

The worst-case data volume can still scale with the combined length of the samples, and the cost of posterior history inference may exceed the cost of reading and comparing their sequences. A feasible staged design would reuse stable variant and annotation indexes, infer local shared factors when supported, and retain sample-specific uncertainty. In a changing cohort, incremental updates should modify only affected paths and factors where possible, with consistency checks for reference changes and coordinate liftover. Structural-variant cycles, duplications, repeats, and ambiguous alignments require careful bookkeeping; the proposed scheme is a research architecture, not an already validated multi-genome algorithm.

14 Direct Topological Analysis on Data-Informed Genomic Two-Complexes

The physically ambitious route developed in Sections 8–9 asks whether an interacting action generates an appropriate complex and a collective limit. A separate, presently more accessible route starts with *observed* genomic or chromosome-organization data, constructs a finite cell complex by explicit rules, and computes its topology without specifying a microscopic action S or solving field equations. This is an analysis of a *data-informed relational representation*; by itself it is neither evidence that the genome is a quantum spacetime nor a derivation of its chemical origin.

14.1 A Construction with Biologically Distinguished Cells

Let $D = (w, \mathcal{A}, C, \mathcal{M}, e)$ denote an observed nucleotide sequence or pangenome paths w , genomic annotations $\mathcal{A}$, a condition-dependent contact matrix or edge set C , optional direct multi-locus observations $\mathcal{M}$, and experimental metadata e . At resolution ℓ , let V_ℓ be a set of aligned genomic loci or bins. A candidate *data-informed genomic complex* is

$$\mathcal{K}_\ell(D) = (V_\ell, E_\ell^{\text{cont}} \cup E_\ell^{\text{prox}}, F_\ell, \omega_\ell, a_\ell), \quad (38)$$

where E_ℓ^{cont} are oriented contour edges, E_ℓ^{prox} encode condition-specific contacts or other specified relations, F_ℓ consists of attached two-cells, ω_ℓ stores confidence and physical/experimental weights, and a_ℓ stores molecular attributes. The cellular boundary maps $B_2 : C_2 \rightarrow C_1$ and $B_1 : C_1 \rightarrow C_0$ must be defined by these cells and satisfy $B_1 B_2 = 0$. They may be stored sparsely. The ordinary sequence is an attribute on an oriented contour, or a compatible collection of sample-specific paths, rather than a complete topological description.

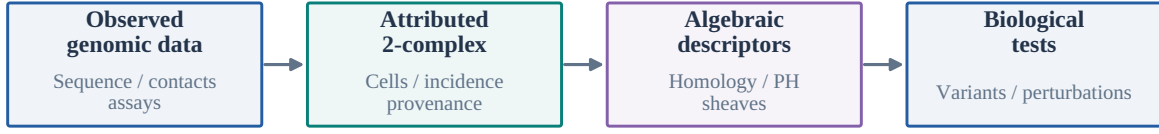
The interpretation of F_ℓ is consequential. At least three admissible but *non-equivalent* designs deserve investigation: (i) *flag faces*, in which every triangle of a contact-threshold graph is filled by definition; (ii) *observational faces*, allowed only when a specified multiway assay supports an appropriate higher-order relation; and (iii) *process faces*, encoding explicitly posited elementary changes in a cellular history. Flag faces express pairwise clique closure and do not prove simultaneous three-locus interaction. Direct multiway techniques, including SPRITE and GAM, have different measurement sensitivities from Hi-C (Fiorillo et al., 2021). Process faces require an event semantics and should not be inferred from static Hi-C triangles without further evidence. The three constructions must carry distinct provenance labels and cannot be merged merely because they share the same incidence matrix shape.

14.2 Filtered Complexes and Measurable Invariants

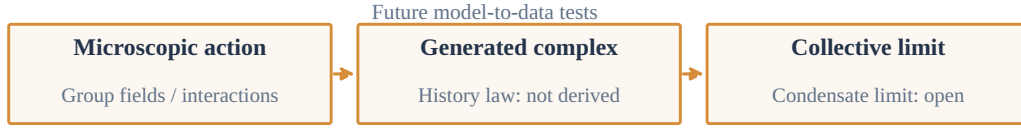
For one particular, fully specified construction, normalize contact scores p_{ij} within comparable data and define

$$E_\lambda = E_\ell^{\text{cont}} \cup \{(i, j) : p_{ij} \geq \lambda\}, \quad \mathcal{K}_\lambda = \text{skel}_2 \text{Flag}(V_\ell, E_\lambda). \quad (39)$$

A Direct genomic topology from observations



B Microscopic generative-physics derivation (not yet established)



Solid arrows: explicit prospective analysis stages. Dashed arrows: research hypotheses, not computational results.

Figure 13: Two complementary routes from genomic observations or generative dynamics. The direct analytical route (upper branch) builds a sparse attributed two-complex using an explicit face rule, computes topological and algebraic descriptors, and evaluates biological associations. The field-theory route (lower branch) seeks a dynamical derivation of complexes and collective limits; its dashed connections mark open theoretical tasks. Data-informed topology can be pursued without completing the lower branch.

When the threshold decreases, the graphs and their flag complexes are nested. Equivalently, a filtration parameter $t = -\lambda$ increases. An explicit persistent homology computation then identifies birth and death of classes in $H_k(\mathcal{K}_\lambda; \mathbb{F})$; an H_1 class is an unfilled algebraic cycle in the *chosen complex*, not automatically a physically observed chromatin loop. A different contact-to-distance transformation changes the filtration and therefore the result. Experimental normalization, missing bins, sequencing depth, genomic-distance biases, batch effects, and cutoff selection must be recorded and controlled.

For a finite two-dimensional cell complex with n_0, n_1, n_2 cells and boundary matrices B_1, B_2 over a field $\mathbb{F}$, elementary rank-nullity yields

$$\beta_0 = n_0 - \text{rank } B_1, \quad \beta_1 = n_1 - \text{rank } B_1 - \text{rank } B_2, \quad \beta_2 = n_2 - \text{rank } B_2. \quad (40)$$

The last identity presupposes that $C_3 = 0$. If a three-dimensional filling is also included, $\beta_2 = n_2 - \text{rank } B_2 - \text{rank } B_3$. Results over $\mathbb{F}_2$ do not retain orientation signs, while real-valued Hodge operators and integral torsion questions may require other coefficients. Consequently, the coefficient system is part of the model specification.

Proposition 14.1 (The fixed support cannot distinguish distinct nucleotide words topologically). *Let $\mathcal{K}$ be a fixed finite cell complex and a, a' two assignments of nucleotides or annotations to its cells that do not modify its incidence maps. Then $H_k(\mathcal{K}; \mathbb{F})$, its Betti numbers, and its unweighted cellular Laplacian spectra are identical under a and a' for all k .*

Proof. The chain groups and boundary operators of $\mathcal{K}$ are unchanged by attribute assignments. Homology and the unweighted Hodge operators depend on those operators alone. $\square$

This elementary obstruction applies, in particular, to nucleotide-labelled versions of the fixed product cylinder $C_N \times I_T$ in Section 3. Comparative genomic topology therefore requires either a data-dependent construction or a genuine *attributed* invariant (for example a specified weight-dependent Laplacian or sheaf), and not merely alternative nucleotide labels on one unaltered cylinder.

14.3 Two-Dimensional Truncation and Artificial Homology

The phrase “genomic two-complex” can conceal an important construction artifact. The two-skeleton of a tetrahedron contains all four triangular faces but no filling three-cell, and therefore has $\beta_2 = 1$ over a field; the full tetrahedron has $\beta_2 = 0$. The apparent two-cycle arises from omitting an admissible higher-dimensional filling. Similarly, an H_2 signal in a contact-derived two-skeleton need not be a biological cavity. We should report whether the object is a genuine two-dimensional CW model with justified faces, a 2-skeleton of a higher-dimensional clique complex, or a truncated approximation whose higher simplices have been intentionally discarded. Computing persistence in dimension two for a full Rips filtration may require information about three-simplices to determine which two-dimensional cycles are killed (Aggarwal and Periwai, 2024).

The definition of a face also affects complexity. Even with sparse vertices, the number of triangles can grow rapidly in dense contact neighborhoods. Exact rank computations and persistent matrix reductions can become expensive despite sparse boundary storage. A realistic chromosome-scale pipeline should construct high-confidence complexes incrementally, cap local density only with an explicit approximation policy, and report the number of cells, nonzero incidences, coefficients, filtration range, time, and peak memory. Sparse homology algorithms already provide an important feasibility precedent: Aggarwal and Periwai (2024) reported persistent homology for a human Hi-C-derived point cloud with approximately three million points in under five minutes with under five gigabytes of memory *for their particular sparse filtration and benchmark*. This is not a performance prediction for an arbitrary genome-labelled CW model. The newer topoHiC work also uses persistent homology to call chromatin loops and TADs (Chiang and Wu, 2026); simply adding PH to Hi-C data is therefore not a new contribution by itself.

14.4 Sparse Storage and the Cost of Higher-Order Cells

Let $n = |V|$, $m = |E|$ and $t = |F|$ for a triangular two-complex. The nonzero incidence count is exactly $2m + 3t$ for ordinary edges and oriented triangular faces, up to repeated-cell conventions; a sparse representation therefore uses $O(n + m + t)$ structural storage in addition to annotations. In a flag construction, the triangle count can be written

$$t = \frac{1}{3} \sum_{\{i,j\} \in E} |N(i) \cap N(j)|, \quad (41)$$

where $N(i)$ is the graph neighborhood. In the worst case $t = \binom{n}{3}$, despite sparse storage of each triangle’s individual incidence. If a full clique filtration is needed to kill degree-two cycles, the number of tetrahedra can be as large as $\binom{n}{4}$. A sparse algorithm avoids storing all such objects when its input geometry and filtration permit implicit enumeration, but it cannot guarantee favorable performance on arbitrary dense or adversarial instances. Runtime also depends on the algebraic reduction order and the complexity of identifying representatives, not solely on the input cell count.

At fixed resolution and fixed observed edges, direct topological descriptors can thus be computed without evaluating a physical path integral or integrating a GFT field equation. This is a materially narrower task, not an automatic reduction of every genome-scale inference problem to linear complexity. Recommended reporting quantities include n, m, t , the number of processed higher-dimensional simplices, the sparsity and coefficient type of each boundary matrix, threshold and normalization rules, wall-clock time, peak memory and the fraction of unresolved or discarded interactions. Since the supplied manuscript does not report such tests on genomic cohorts, all chromosome-scale figures discussed here are external feasibility precedents or proposed targets.

14.5 Localized Cycles, Relative Topology, and Hodge Features

A global Betti vector discards the locations of the loci responsible for the classes. For biological interpretation, a useful output should include selected cycle representatives, their genomic intervals, confidence, contacts and associated annotations, while acknowledging that representatives of one homology class are not unique. Methods for selecting tight persistent representatives have already been demonstrated on genomic data (Aggarwal and Periwai, 2023).

If $A \subseteq \mathcal{K}$ is a specified contour or conserved reference subcomplex, relative homology is defined through $C_\bullet(\mathcal{K}, A) = C_\bullet(\mathcal{K})/C_\bullet(A)$ and its induced differential; it measures classes *relative to the chosen subcomplex*. There is no canonical biological “relative boundary” until A has been fixed. For real cellular chains with a chosen inner product, the unweighted Hodge operators are

$$L_0 = B_1 B_1^\top, \quad L_1 = B_1^\top B_1 + B_2 B_2^\top, \quad L_2 = B_2^\top B_2, \quad (42)$$

and $\dim \ker L_k = \beta_k(\mathcal{K}; \mathbb{R})$. Weighted Laplacians require specified positive inner products on chain spaces and should not be treated as having an invariant spectrum independent of cell weighting. These operators offer sparse spectral descriptors; their biological meaning must be established through perturbations, controls, and external measurements.

15 Stability and Statistical Reliability of Genomic Topological Comparisons

The preceding construction makes topological summaries computable, but the meaning of differences between two observed complexes depends on how contact weights, missing loci, assay variation, and cell-attachment rules enter the construction. A topological descriptor can be mathematically well defined while remaining unstable under an ill-specified data-processing procedure. This section identifies the conditions under which persistent features are stable, separates score perturbations from changes in the underlying cellular support, and states a statistical validation protocol for future genomic applications. The results are mathematical consequences of a fixed filtration model; no biological confidence interval or chromosome-scale experimental result is claimed.

15.1 Score-Based Filtrations and a Perturbation Bound

Fix a finite vertex set V , a set E_0 of mandatory contour edges, and a common candidate set E_c of potential contact edges. Let $p : E_c \rightarrow [0, 1]$ be a calibrated contact score, with larger values corresponding to earlier appearance in a superlevel filtration. For a finite simplicial complex, define the entrance time of vertices and mandatory contour edges to be -1 and the entrance time of a candidate contact edge e to be $-p_e$. For every triangular face introduced by the two-skeleton flag rule, set its entrance time equal to the latest entrance of its boundary edges. In symbols,

$$f_p(v) = -1, \quad f_p(e) = \begin{cases} -1, & e \in E_0, \\ -p_e, & e \in E_c, \end{cases} \quad f_p([u, v, z]) = \max\{f_p(uv), f_p(vz), f_p(zu)\}. \quad (43)$$

The sublevel sets $\mathcal{K}_p(t) = \{\sigma : f_p(\sigma) \leq t\}$, with only simplices whose entire boundary is included, form an increasing two-dimensional filtration. This is exactly the flag-filtration convention of Eq. (39) after the substitution $t = -\lambda$. A fixed candidate-edge universe is essential: an unmeasured contact is not automatically an observed contact with score zero. The construction can be carried out over any coefficient field, although the examples below use $\mathbb{F}_2$.

Proposition 15.1 (Stability of a fixed-support genomic flag filtration). *Let $p, p' : E_c \rightarrow [0, 1]$ be two score assignments on the same V, E_0, E_c , with the same simplicial face rule. If $\|p - p'\|_\infty \leq \varepsilon$,*

then the induced persistent homology modules satisfy an ε -interleaving. For every dimension $k \in \{0, 1, 2\}$, their persistence diagrams, including matching essential intervals when present, satisfy

$$d_B(D_k(\mathcal{K}_p), D_k(\mathcal{K}_{p'})) \leq \varepsilon, \quad (44)$$

where d_B is the bottleneck distance on the common filtration parameter scale.

Proof. For each contact edge, $|f_p(e) - f_{p'}(e)| = |p_e - p'_e| \leq \varepsilon$; entrance times for fixed vertices and contour edges agree. The maximum of a finite list is 1-Lipschitz in the sup norm, hence every triangular face satisfies $|f_p(\sigma) - f_{p'}(\sigma)| \leq \varepsilon$. Thus every simplex of $\mathcal{K}_p(t)$ belongs to $\mathcal{K}_{p'}(t + \varepsilon)$, and conversely. The two inclusion families commute with the filtration structure maps, giving an ε -interleaving. The bottleneck stability theorem for finite persistence modules then implies the bottleneck inequality (Cohen-Steiner et al., 2007). Both filtrations terminate in the same finite complex, so essential classes have matching multiplicities and do not obstruct the bound. $\square$

This is a specialized consequence of existing persistent-homology stability theory, not a novel general stability theorem. It nevertheless gives a practical specification: a claimed bound on the topology of a contact-derived complex requires a *uniform bound on the actual score transformation on a shared support*. The inequality does not promise that Betti numbers at a single, sharply chosen threshold are unchanged. A contact near the threshold may enter on opposite sides and create or fill a cycle; the persistence diagram records how the corresponding interval changes. Furthermore, a finite persistence interval of length greater than 2ε cannot be matched to the diagonal within an ε -matching, although its associated genomic interpretation still needs independent evidence.

15.2 Confidence-Controlled Comparisons and the Limits of Stability

Assume that $\hat{p}$ estimates an unobserved, well-defined score vector p^* on the same set of candidate edges and that a measurement model supplies a simultaneous error statement

$$\mathbb{P}(\|\hat{p} - p^*\|_\infty \leq \varepsilon) \geq 1 - \alpha. \quad (45)$$

This is an assumption to be validated, not a confidence bound furnished automatically by persistent homology. Proposition 15.1 yields the following directly.

Corollary 15.2 (Diagram uncertainty inherited from simultaneous score bounds). *Under Eq. (45),*

$$\mathbb{P}(d_B(D_k(\mathcal{K}_{\hat{p}}), D_k(\mathcal{K}_{p^*})) \leq \varepsilon) \geq 1 - \alpha. \quad (46)$$

For two samples A, B registered to a common candidate support, suppose their score bounds hold with error levels $\varepsilon_A, \varepsilon_B$ and failure probabilities α_A, α_B . Then, with probability at least $1 - \alpha_A - \alpha_B$,

$$|d_B(\hat{D}_{k,A}, \hat{D}_{k,B}) - d_B(D_{k,A}^*, D_{k,B}^*)| \leq \varepsilon_A + \varepsilon_B. \quad (47)$$

Proof. The first claim follows by applying Proposition 15.1 on the event in Eq. (45). For the second, intersect the two score-control events, use the union bound for their probability, and apply the triangle inequality for the bottleneck metric twice. $\square$

The latter statement bounds distortion of a *topological comparison*, rather than providing a test of equal gene regulation or chromatin function. Even an observed diagram distance exceeding $\varepsilon_A + \varepsilon_B$ only distinguishes the chosen score-based filtrations under the declared uncertainty model. It neither identifies which genomic loci cause the difference nor establishes a physical mechanism. Localized generators, functional annotations, and interventions remain separate steps.

There are three important exclusions from this stability guarantee. First, a change in the vertex universe or missing-data mask does not generally admit an elementwise sup-norm comparison;

coordinate matching, masked subcomplexes, or a different metric for changing supports must be defined. Second, when the cell-attachment rule differs, for instance between clique-derived triangles and directly measured three-way interactions, the two filtrations need not interleave under a bound on pairwise scores alone. Third, nonlinear normalization may amplify small raw-count changes, and a score scale estimated separately in each sample need not be metrically comparable. Cross-sample normalization and batch effects in Hi-C have been documented explicitly (Fletez-Brant et al., 2024); the existence of a general topological stability theorem cannot substitute for a suitable experimental normalization model.

15.3 Experimental Replicates, Cellular Heterogeneity, and Filtration Uncertainty

A proposed empirical implementation should use replicates and experimental metadata to establish the score-error model before interpreting barcode differences. For bulk Hi-C, contact events are affected by read depth, mapping bias, genomic separation, copy-number changes, and cell-population heterogeneity. Treating all observed read pairs or all edges as independent samples can understate uncertainty because a cell, sample, library, or genomic neighborhood may contribute correlated observations. Comparisons should therefore use resampling units reflecting the biological and technical design, such as independent biological replicates or appropriately modeled cell-level data. Hi-C quality and replicate metrics, including distance-stratified approaches, already offer essential baselines (Yardimci et al., 2019; Fletez-Brant et al., 2024).

A prespecified validation design should distinguish four kinds of variation. *Measurement variation* concerns contact-score uncertainty conditional on fixed biological structure. *Support variation* concerns missing loci, assembly ambiguity, or changed correspondence between genomes. *Construction variation* concerns whether a face is a flag closure, a separately measured multiway relation, or an assumed process cell. *Biological variation* concerns actual differences between individuals, cell types, interventions, or timepoints. Bootstrap confidence procedures can help characterize some of these components, but resampling does not automatically separate them. Where a joint score bound of the form (45) cannot be justified, we should report empirical variability of topological summaries without presenting Eq. (46) as an inferential guarantee.

For multiple testing across chromosomes, genomic windows, homology degrees, filtration choices, and annotated generators, discovery claims require explicit control of the testing family. A useful future analysis would hold out biological replicates, estimate feature stability on training data, freeze the cell and score rules, and evaluate between-condition differences on held-out material. Alongside bottleneck distance, the analysis should report structural overlap, the genomic positions supporting selected cycles, reproducibility of face provenance, and independently assessed functional or perturbational relevance. A statistically distinguishable persistence diagram is an intermediate result, not a biological endpoint.

15.4 A Resource-Aware Reliability Protocol for Large Complexes

The stability bound is computationally inexpensive to state but not necessarily cheap to operationalize. Computing $\|p - p'\|_\infty$ across M aligned candidate edges requires $O(M)$ comparisons once the score vectors are available. Constructing every triangular face may be substantially more expensive, with the number of candidate triangles as large as $\binom{|V|}{3}$ in a dense flag model, before persistent-homology reduction. Accordingly, candidate-contact screening, sparse adjacency indexing, bounded local neighborhoods, and biologically motivated face selection should precede large-scale persistence computations. Sparsification may change the resulting complex and invalidate a claim of preserving topology unless accompanied by its own approximation or interleaving bound.

The minimum report for a future genome-wide result should specify genomic resolution, fixed

and candidate edges, the exact score normalization and score domain, any missing-data policy, coefficients for homology, face provenance, filtration convention, the number of cells by dimension, sparse memory use, diagram computation time, and how replicates and uncertain correspondences were handled. The direct genomic-complex route can then be assessed on its own terms: stable features and reproducible biological associations under explicit uncertainty, rather than inferred evidence for a microscopic group-field action. This is the operational bridge between the present mathematical construction and a credible comparative-genomics method.

16 Multiresolution Correspondence and the Preservation of Genomic Topological Information

The direct topological program in Sections 14 and 15 presumes that contact-derived complexes can be represented at more than one genomic resolution. Yet an aggregation of sequence bins is not, by itself, a topology-preserving map. This section supplies the missing correspondence between cellular representations: explicit maps of chains, conditions for preservation of homology and persistence, an exact finite example distinguishing safe reduction from destructive collapse, and tests that may be implemented before biological interpretation. It is independent of the microscopic group-field action in Section 8; none of the claims below establishes a physical condensate or supplies an empirically calibrated genome-scale algorithm.

16.1 Cellular Maps Between Genomic Resolutions

Fix a biological sample, experimental condition, and construction rule for edges and higher-order cells. Let $\mathcal{K}^f$ and $\mathcal{K}^c$ be finite fine- and coarse-resolution complexes, with chosen coefficient ring R (initially $R = \mathbb{Z}$) and cellular boundary operators B_k^f, B_k^c . A *resolution correspondence* is a family of linear maps

$$F_k : C_k(\mathcal{K}^f; R) \longrightarrow C_k(\mathcal{K}^c; R), \quad B_k^c F_k = F_{k-1} B_k^f, \quad k = 1, 2. \quad (48)$$

The equalities express cellular boundary compatibility and are directly auditable as sparse matrix identities. A cellular map of CW complexes induces such a chain map, although an arbitrary assignment of fine bins to coarse bins need not do so (Hatcher, 2002). In particular, defining F_0 by an interval correspondence does not determine a valid F_1 on contact edges or a valid F_2 on relational faces. Those maps must be constructed, or the correspondence rejected.

For a designated genomic contour Γ^f , preserving the *ordered contour* imposes an additional condition: the map should respect chromosome identity, orientation, and the relevant endpoints or cyclic orientation. Depending on the task, the desired requirement may be $F_1(\Gamma^f) = \Gamma^c$ as cellular chains, or a relative map of pairs $F : (\mathcal{K}^f, A^f) \rightarrow (\mathcal{K}^c, A^c)$, where A consists of designated boundary subcomplexes. For linear chromosomes, endpoint correspondence matters; for circular chromosomes, degree on the contour cycle matters. Structural variations, ambiguous repeats, and unaligned regions can prevent the existence of a unique biologically meaningful correspondence.

Proposition 16.1 (Homological functoriality of resolution correspondences). *If the maps F_k satisfy Eq. (48), they induce homomorphisms*

$$H_k(F) : H_k(\mathcal{K}^f; R) \longrightarrow H_k(\mathcal{K}^c; R), \quad [z] \longmapsto [F_k z]. \quad (49)$$

If there is a reverse chain map G such that $GF - \text{id}$ and $FG - \text{id}$ are chain-homotopic to zero, then $H_k(F)$ is an isomorphism for every k .

Proof. For a cycle z , $B^c F z = F B^f z = 0$, so $F z$ is a cycle; for a boundary $z = B^f y$, $F z = B^c F y$ is a boundary. The induced map is therefore independent of the representative. If $GF - \text{id} = BH + HB$ and $FG - \text{id} = BH' + H'B$ for degree-raising operators H, H' , then on cycles these differences are boundaries. Hence $H_k(G)H_k(F)$ and $H_k(F)H_k(G)$ are the respective identity maps. $\square$

This proposition separates two frequently conflated properties. Equation (48) guarantees a *well-defined comparison*, while a chain-homotopy equivalence guarantees preservation of ordinary homology. Neither condition guarantees preservation of contact strengths, molecular labels, unweighted or weighted Laplacian eigenvalues, or biological function. Those require independent compatibility assumptions and tests.

16.2 An Exact Cellular Coarse-Graining with Surviving and Lost Cycles

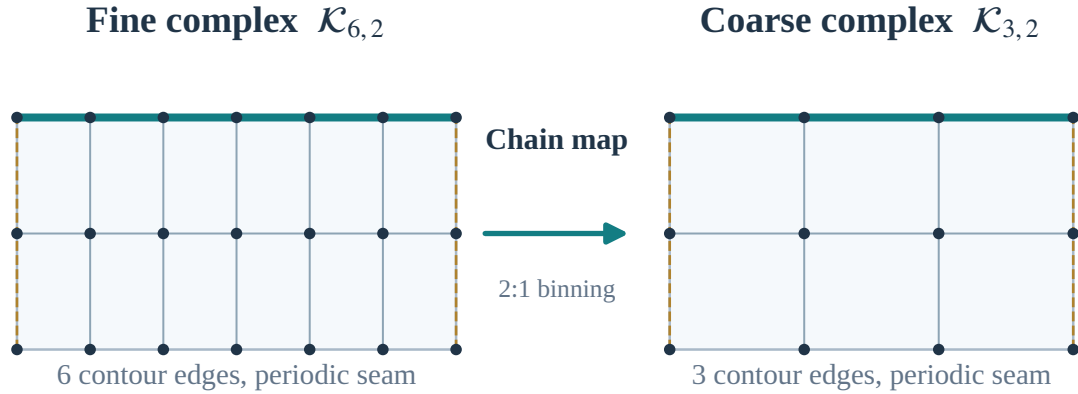
A small construction shows why the distinction is operationally important. Let $\mathcal{K}_{N,T} = C_N \times I_T$ be the product cylinder defined in Section 3, and consider $N = 6$, $T = 2$. Its contour has six edges at each history slice. A coarse counterpart $\mathcal{K}_{3,2}$ has three per slice. For $i = 0, \dots, 5$ and $j = 0, 1, 2$, define a map on vertices and horizontal edges by

$$\begin{aligned} F_0(v_i^j) &= \bar{v}_{[i/2]}^j, & F_1(a_{2r}^j) &= 0, & F_1(a_{2r+1}^j) &= \bar{a}_r^j, \\ F_1(b_i^j) &= \bar{b}_{[i/2]}^j, & F_2(f_{2r}^j) &= 0, & F_2(f_{2r+1}^j) &= \bar{f}_r^j, \end{aligned} \quad (50)$$

where $r = 0, 1, 2$ and the vertical-edge and face formulas apply for $j = 0, 1$. Bars denote coarse cells, and cyclic indices are taken modulo the relevant contour size. Every second fine horizontal edge and the corresponding square face collapse to a lower-dimensional cell. The other edges and faces map with positive orientation. Direct substitution into Eq. (3) verifies the chain relations in Eq. (48). More importantly, the terminal generator is preserved:

$$F_1 \left(\sum_{i=0}^5 a_i^2 \right) = \sum_{r=0}^2 \bar{a}_r^2. \quad (51)$$

The two cylinders have Betti numbers $(1, 1, 0)$ and the map has degree $+1$ on the one-dimensional homology generator. Consequently, it induces an isomorphism on H_1 over $\mathbb{Z}$.



Preserving map: $H_1(\mathcal{K}_{6,2}) \cong H_1(\mathcal{K}_{3,2}) \cong \mathbb{Z}$

Cycle-collapsing map: The contour contracts to a point: $H_1(F) = 0$.

A valid chain map preserves incidence identities; topology preservation requires an induced-map test.

Figure 14: An explicit six-to-three-bin cellular correspondence on product cylinders. The top contours represent the same generator of first homology, although half of the fine contour edges map to zero and the remaining edges map to the coarse edges. The lower statement concerns a different, equally valid chain map from the fine cylinder to an interval, which kills the contour cycle. In both cases the algebraic boundary identity holds; only the first map preserves the selected H_1 information. The illustration concerns constructed supports, not measured chromosome topology.

By contrast, there is a chain map $Q : \mathcal{K}_{6,2} \rightarrow I_2$ sending every vertex in a given slice to one interval vertex, every horizontal edge to zero, every vertical edge to its interval edge, and every face to zero. It still satisfies $B_1^f Q_1 = Q_0 B_1^f$ and $Q_1 B_2^f = 0$. Nevertheless, $Q_1(\Gamma_2) = 0$ and $H_1(I_2; \mathbb{Z}) = 0$. Thus the original cycle is destroyed without violating a single chain identity.

These claims can be checked on integer matrices without simulation or fitting. For $\mathcal{K}_{6,2}$ the incidence dimensions are $B_1 \in \mathbb{Z}^{18 \times 30}$ and $B_2 \in \mathbb{Z}^{30 \times 12}$; for $\mathcal{K}_{3,2}$ they are $\bar{B}_1 \in \mathbb{Z}^{9 \times 15}$ and $\bar{B}_2 \in \mathbb{Z}^{15 \times 6}$. Independent exact-rank calculations give $(\beta_0, \beta_1, \beta_2) = (1, 1, 0)$ for each cylinder, while direct matrix products check every boundary and correspondence identity. This is an exact finite verification of the construction, *not* a quantitative benchmark on real genomic measurements.

16.3 Quantifying Topological Information Retained by Aggregation

For a field $\mathbb{F}$ and a specified chain map F , let $r_k(F) = \text{rank}_{\mathbb{F}} H_k(F)$ and define the two nonnegative integers

$$L_k(F) = \beta_k^f - r_k(F), \quad E_k(F) = \beta_k^c - r_k(F). \quad (52)$$

The first counts dimensions of fine-resolution homology classes annihilated by aggregation; the second counts dimensions of coarse-resolution classes not reached by fine-resolution classes. For a selected degree k , preservation requires $L_k = E_k = 0$; in this case $H_k(F)$ is an isomorphism over $\mathbb{F}$. One may instead focus on a prescribed subspace generated by experimentally localized classes and report its image rank. The numerical values depend on the coefficient field and the chosen chain map. Over $\mathbb{Z}$, torsion and multiplication degrees must be examined separately, since Betti ranks do not characterize integral homology maps.

A barcode, Betti vector, or cohomology ring alone does not tell us which individual class in one resolution corresponds to a class in another. The induced map provides this additional information. Likewise, equal Betti numbers are insufficient to infer $L_k = E_k = 0$: a map can annihilate a generator while both source and target have nonzero H_k . For biologically annotated cycles, preservation should additionally require correspondence of their supporting loci, edge semantics, and data provenance. A cycle whose representative changes under aggregation may still represent the same homology class, whereas a biologically meaningful contact can disappear even if the abstract homology class persists.

Sequence features require a separate aggregation operator. Let $p_i \in \Delta^3$ encode a four-state composition vector or a probability distribution on nucleotides. A coarse-bin summary may be defined by $\bar{p}_B = \sum_{i \in B} \alpha_{Bi} p_i$, with nonnegative coefficients summing to one for each bin. This operation differs from the cellular chain map F_1 : a fine edge collapsed as a chain can still contribute to a coarse-bin composition estimate. The nonlinear discrete readout $B(\Phi)$ generally does not commute with averaging, as discussed in Section 10. Preservation of cellular homology is therefore neither necessary nor sufficient for preservation of sequence identity, motif order, or predicted regulation.

16.4 Filtered Resolution Maps and Persistence Guarantees

The contact-score perturbation theorem of Section 15 compares two filtrations on a *shared* cell support. Coarse-versus-fine comparisons require a different guarantee. Let $\{\mathcal{K}_t^f\}_{t \in \mathbb{R}}$ and $\{\mathcal{K}_t^c\}_{t \in \mathbb{R}}$ be finite-type filtrations, with increasing parameter t and field coefficients. Suppose there are filtration-compatible chain maps

$$F^t : C_\bullet(\mathcal{K}_t^f) \longrightarrow C_\bullet(\mathcal{K}_{t+\varepsilon}^c), \quad G^t : C_\bullet(\mathcal{K}_t^c) \longrightarrow C_\bullet(\mathcal{K}_{t+\varepsilon}^f), \quad (53)$$

whose two composites are chain-homotopic to the respective filtration inclusion maps shifted by 2ε , with the homotopies chosen compatibly with persistence structure maps. These are stronger requirements than assigning the same resolution label to two independently constructed complexes.

Proposition 16.2 (Conditional persistence preservation across resolutions). *Under the hypotheses of Eq. (53), the degree- k persistent homology modules of the fine and coarse filtrations are ε -interleaved. If the resulting modules admit ordinary persistence diagrams (in particular, for finite filtrations), then*

$$d_B(D_k^f, D_k^c) \leq \varepsilon. \quad (54)$$

Proof. Applying homology to the filtered chain maps gives persistence-module morphisms with shift ε . Chain-homotopic composites induce the same homology maps as the inclusion morphisms at shift 2ε , so the defining interleaving identities hold. The standard algebraic stability/isometry result for one-parameter persistence modules then yields the bottleneck bound (Chazal et al., 2016). $\square$

A topological simplification algorithm must *construct and verify* the stated maps before invoking this result. Generic bin merging, threshold normalization, contact imputation, or graph clustering provides no such guarantee. Even when Eq. (54) holds, short-lived features near the resolution error may be indistinguishable; anatomical or functional annotations need separate uncertainty assessment. The bound addresses persistence modules, not the recovery of unique geometric representatives or biochemical mechanisms.

16.5 Cross-Genome Compatibility and Computational Requirements

For two genomes or conditions A and B , a matched fine-resolution relation $T^f : C_\bullet(\mathcal{K}_A^f) \rightarrow C_\bullet(\mathcal{K}_B^f)$ and a coarse relation T^c need not be consistent with the aggregation maps F_A, F_B . A stronger notion of cross-genome comparability asks whether

$$F_B T^f \simeq T^c F_A \quad (55)$$

as *chain maps*, where $\simeq$ denotes chain homotopy. When this identity holds, the induced homology maps form a commuting square, so cross-genome comparisons are compatible with the selected resolution change. If no biologically justified maps T^f, T^c exist, the comparison should be reported as a distance between separately constructed descriptors, not as a transformation of identified topological classes. Pangenome graphs, syntenic intervals, chromosome-specific orientation, and mapping uncertainty provide candidate correspondence information; none supplies Eq. (55) automatically.

For sparse chain matrices and locally supported F_k , storage scales with $O(\sum_k \text{nnz}(B_k) + \sum_k \text{nnz}(F_k))$, rather than the square of the number of genomic bins. Checking Eq. (48) entails sparse matrix products; the true cost depends on their intermediate fill-in. Computing induced image ranks or persistent maps is generally more expensive, and arbitrarily dense contact graphs can generate combinatorially many flag faces before any homology calculation begins. Practical multi-resolution experiments should therefore fix cell-admissibility rules, bound higher-order cell generation, preserve chromosome and sample identifiers, and record each many-to-one bin assignment with its uncertainty.

A defensible evaluation would construct at least two resolutions of the *same* genomic observations, validate the cellular maps and any designated boundary pairs, calculate L_k, E_k and class-level persistence correspondence, and compare these measures with a simple contact-graph baseline. It would then test whether the retained features predict held-out perturbation effects or reproduce independently measured contacts. Failure of chain-map consistency, large unexplained class loss, or disappearance of biological associations under small reasonable changes in resolution would count against the proposed construction. The core methodological conclusion is precise: *multiscale genomic topology requires maps between representations, not merely multiple representations at different bin sizes.*

17 Structural Variation and Partial Cellular Correspondences Between Genomes

Direct comparison of genomic complexes becomes substantially more difficult when the chromosomes differ by inversion, duplication, deletion, insertion, or translocation. Such changes may alter sequence order, local copy number, and the placement of physical contacts simultaneously. The resolution maps of Section 16 operate between representations of a selected biological sample; they need not supply a map between two different genomes. The aim of this section is to formulate partial and ambiguous genomic correspondences as explicit cellular relations, characterize which topological classes can be jointly supported, and specify the limits of what comparative topology can infer from an alignment. The construction is intended for observed or reconstructed complexes and does not presuppose the microscopic action of Section 8.

17.1 Cell Types, Orientation, and Partial Genomic Alignment

Consider finite attributed cellular complexes $\mathcal{K}_A$ and $\mathcal{K}_B$ for two haplotypes, assemblies, or experimentally observed genomic states. A correspondence must specify more than a pair of index lists. Each matched cell requires an underlying coordinate reference, its sequence orientation, copy or paralog identity where identifiable, chromosomal provenance, a biological relationship type, and an uncertainty annotation. Contour edges, chromatin-contact edges, and generative-history edges are distinct types; a proposed map need not preserve a contact edge simply because it preserves its two endpoint loci. Likewise, a two-cell attributed to a verified multiway contact should not be matched to a flag-complex face merely because the face boundaries coincide combinatorially.

A simple genome-to-genome chain map $T_\bullet : C_\bullet(\mathcal{K}_A; \mathbb{F}) \rightarrow C_\bullet(\mathcal{K}_B; \mathbb{F})$ would need to satisfy

$$\partial_k^B T_k = T_{k-1} \partial_k^A, \quad k = 1, 2. \quad (56)$$

A structural-variant alignment generally is *not* such a total cellular map. In particular, a homology statement for the endpoints of an ancestral edge does not imply that the corresponding loci remain adjacent after a rearrangement. One can always join two vertices by some path in a connected target graph, but using that path as the image of an ancestral edge may manufacture a spurious claim of conserved adjacency. Thus the algebraic possibility of satisfying Eq. (56) is weaker than its biological admissibility. Existing pangenome graphs address many of the underlying alignment and structural-variant problems, with nontrivial differences in their handling of inversions and duplicated segments (Hickey et al., 2023; Salehi Nowbandegani et al., 2026).

17.2 Common Cellular Supports and Correspondence Spans

Instead of requiring a total transformation from $\mathcal{K}_A$ to $\mathcal{K}_B$, select a *common evidence complex* $\mathcal{Z}$ carrying only cells for which a correspondence is admissible under an explicit alignment and cell-type policy. Over a field $\mathbb{F}$, define cellular chain maps

$$C_\bullet(\mathcal{K}_A; \mathbb{F}) \xleftarrow{f_\bullet} C_\bullet(\mathcal{Z}; \mathbb{F}) \xrightarrow{g_\bullet} C_\bullet(\mathcal{K}_B; \mathbb{F}), \quad \partial^A f = f \partial^\mathcal{Z}, \quad \partial^B g = g \partial^\mathcal{Z}. \quad (57)$$

Neither f nor g is assumed injective, surjective, or a homotopy equivalence. In the most conservative case $\mathcal{Z}$ is a subcomplex of independently supported homologous contour intervals, with orientation-corrected embeddings in the two genomes. More general chain-level maps may be justified if the intervening paths and cells are themselves supported by data. Sequence or biological labels must be carried by separate maps of attributed data; a chain map does not by itself preserve a nucleotide string or a gene annotation. This formulation also accommodates a matched subcomplex that occupies disjoint intervals on one or both chromosomes.

For a specific inversion, let the conserved interior be the path $a \rightarrow b \rightarrow c$, and the outer anchors be L, R . Genome A has ordered contour L, a, b, c, R , while B has L, c, b, a, R (Fig. 15). Let $e_{ab}^\mathcal{Z}$

Partial cellular correspondence across an inversion

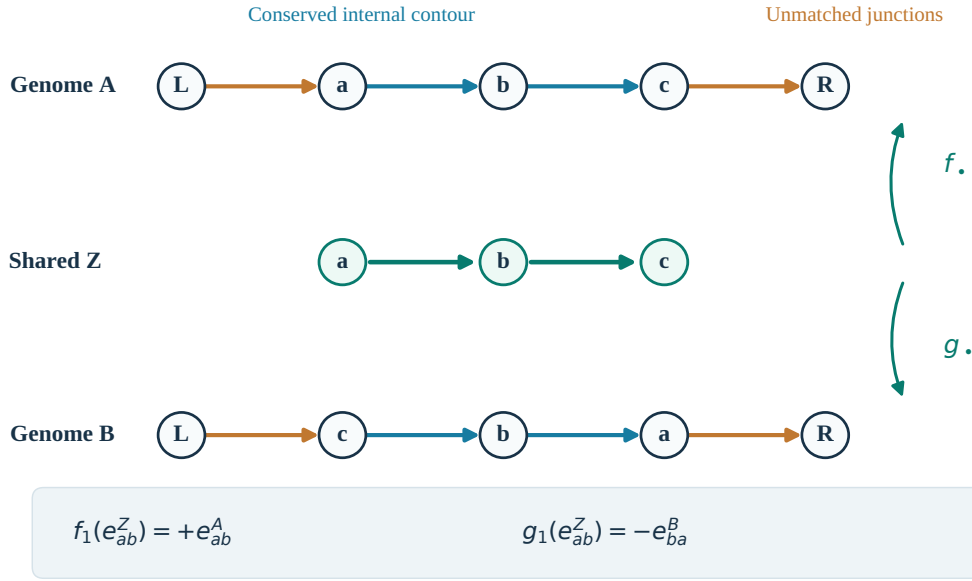


Figure 15: Partial cellular correspondence in a schematic inversion. Genome A contains the oriented contour L, a, b, c, R , whereas genome B contains L, c, b, a, R . The common evidence complex $\mathcal{Z}$ retains only the internal $a-b$ and $b-c$ adjacencies: its maps into B reverse edge orientation, while the variant-specific junctions are excluded. The schematic emphasizes homology and adjacency as distinct assertions. It is not an inferred biological history or an experimentally measured chromatin complex.

denote the oriented edge $a \rightarrow b$ in $\mathcal{Z}$. In A set $f_1(e_{ab}^Z) = e_{ab}^A$; in B, with its contour edges oriented from L toward R , set $g_1(e_{ab}^Z) = -e_{ba}^B$. Then

$$\partial_1^B g_1(e_{ab}^Z) = -(a - b) = b - a = g_0 \partial_1^Z(e_{ab}^Z), \quad (58)$$

where $\partial e_{uv} = v - u$. The same calculation holds for $b \rightarrow c$. The new junctions $L \rightarrow c$ and $a \rightarrow R$ in B are not treated as copies of $L \rightarrow a$ and $c \rightarrow R$ in A. A real DNA inversion also requires correct reverse-complement handling in the separate symbolic readout; signed cellular edges alone do not implement nucleotide complementation.

17.3 Homological Information Jointly Supported by a Correspondence

Set $W_k = H_k(\mathcal{Z}; \mathbb{F})$, $A_k = H_k(\mathcal{K}_A; \mathbb{F})$, $B_k = H_k(\mathcal{K}_B; \mathbb{F})$, and denote the induced maps by f_*, g_* . Their joint correspondence is the linear relation

$$\mathcal{R}_k = \text{im} \begin{pmatrix} f_* \\ g_* \end{pmatrix} \subseteq A_k \oplus B_k. \quad (59)$$

A class in $\mathcal{Z}$ may vanish in either genome, so the dimension of $\mathcal{R}_k$ is not an appropriate count of *shared* surviving classes. The following elementary quantity avoids that mistake.

Definition 17.1 (Jointly supported homology dimension). *For finite-dimensional W_k , the jointly supported dimension of the chosen correspondence is*

$$m_k(f, g) = \dim[W_k / (\ker f_* + \ker g_*)]. \quad (60)$$

It is a property of the specified span, not a complete invariant of two genomes.

Proposition 17.2 (Computable rank formula and bounds). *For the correspondence in Eq. (57),*

$$m_k(f, g) = \text{rank } f_* + \text{rank } g_* - \text{rank} \begin{pmatrix} f_* \\ g_* \end{pmatrix}, \quad 0 \leq m_k(f, g) \leq \min\{\text{rank } f_*, \text{rank } g_*\}. \quad (61)$$

Proof. Write $U = \ker f_*$ and $V = \ker g_*$. Rank-nullity gives $\text{rank } f_* = \dim W_k - \dim U$ and $\text{rank } g_* = \dim W_k - \dim V$. The kernel of the stacked map is $U \cap V$. Consequently, the right-hand side of Eq. (61) equals $\dim W_k - \dim U - \dim V + \dim(U \cap V) = \dim W_k - \dim(U + V)$, proving the identity and the bounds. $\square$

The quotient is an algebraic common signal supported by the selected evidence complex. It does not establish that matched cycles are experimentally homologous, that their localized representatives occupy identical physical positions, or that a structural variant caused the difference. When f_* is an isomorphism, $g_* f_*^{-1}$ supplies a genuine homology map $A_k \rightarrow B_k$; without that condition it is generally invalid to divide by f_* or report a unique transported homology class. If the selected correspondence yields $m_k = 0$, it may reflect disjointly supported classes, insufficient alignment, or genuinely different topology; these explanations require independent information.

17.4 Exact Finite Checks, Copy Ambiguity, and Structural Events

A minimal linear-algebra example makes the limitation concrete. Over $\mathbb{F}_2$, take $W_1 = \mathbb{F}_2^2$ and $A_1 = B_1 = \mathbb{F}_2$. If $f_* = (1 \ 0)$ and $g_* = (1 \ 0)$, then both generators detected by the two target maps are supported by the *same* source class, and Eq. (61) yields $m_1 = 1$. If instead $g_* = (0 \ 1)$, both target groups still have rank one, but their detections are supported on *different* source classes and $m_1 = 0$. This comparison is an algebraic counterexample to equating matched Betti counts with a class correspondence. Exact rank computations for these matrices are given in the accompanying verification notes; no genome-derived homology map was inferred from data.

Duplication requires special treatment. An ancestral interval may align to two descendant copies, while its flanking interactions differ between copies. A choice of only one descendant branch makes a valid *candidate* span, but can erase a second legitimate correspondence. To avoid treating alternatives as a single well-defined map, one may retain a finite family $\{(\mathcal{Z}_j, f^{(j)}, g^{(j)}, \pi_j)\}_j$ with nonnegative weights π_j summing to one, based on sequence evidence and provenance. Reporting the range or weighted distribution of m_k across admissible alternatives expresses alignment uncertainty; it is not a posterior distribution unless the weights arise from a justified probabilistic model. Orthology versus paralogy, shared ancestry versus independently generated contacts, and assembly errors must be separately addressed. Deletions lead to missing supports, while translocations require chromosome-aware anchor maps and may change the connected-component structure of the constructed contact complex.

Inversion changes genomic orientation and breakpoint adjacency; it does not necessarily change the homology of the underlying contour support. A copied interval can change repeat multiplicity without creating a new first-homology class. Conversely, attaching or removing contact edges and two-cells may change H_1 even when the nucleotide order is conserved. These observations caution against treating every structural-variant category as a deterministic operation on Betti numbers. The topology depends jointly on the representation, contact data, and cell-attachment semantics, as established in Sections 14 and 16.

17.5 Filtered Correspondences and Confidence Calibration

For contact-derived filtrations $\{\mathcal{K}_{A,t}\}$ and $\{\mathcal{K}_{B,t}\}$, a single unfiltered alignment does not yield a persistence correspondence. One needs compatible filtered evidence complexes $\{\mathcal{Z}_t\}$ and filtration-compatible maps f^t, g^t whose squares with the inclusion morphisms commute, at least after

passing to homology. This forms a span of persistence modules rather than an ordinary one-directional interleaving; the latter needs additional shifted maps and composition conditions as in Proposition 16.2. Structural events that add and remove cells may instead be studied with a *zigzag* sequence of inclusions when appropriate, using established zigzag persistence theory (Carlsson and de Silva, 2010). Neither construction automatically matches specific cycles to genes or provides a structural-variant likelihood.

Alignment uncertainty and contact-score uncertainty are different. Section 15 bounds diagram variation under score perturbations on a shared support; it does not bound a change in the evidence span itself, nor a face that appears because an alignment selected a different paralog. For each comparison we therefore propose reporting (i) confidence and provenance of matched loci, (ii) strand and copy-number conventions, (iii) all allowed edge and face types, (iv) score stability under fixed correspondence, and (v) variation over credible alignment alternatives. Claims of conservation should survive reasonable alternative correspondences, not merely one favorable alignment.

17.6 Genome-Scale Implementation and Falsifiable Evaluation

A scalable implementation would not materialize a complete cross-product of all cells in two genomes. It would begin with candidate orthologous or syntenic blocks, use existing pangenome paths and structural-variant calls to define oriented partial anchors, then construct locally supported f_k, g_k only on verified cellular supports. Each generated map must pass the sparse identities $\partial^A f = f \partial^Z$ and $\partial^B g = g \partial^Z$ before any homology correspondence is reported. For a fixed pair of sparse complexes, the memory footprint of the matrices is $O(\sum_k \text{nnz}(\partial_k^A) + \text{nnz}(\partial_k^B) + \text{nnz}(f_k) + \text{nnz}(g_k))$; induced homology ranks and filtered correspondence calculations can cost considerably more. Real genome-scale complexity depends on mapping multiplicity, cell counts, fill-in, and persistence reduction, not only the number of nucleotides.

An initial benchmark should compare a matched genome pair with a controlled simulated inversion, duplication, and deletion, followed by independently assembled haplotypes with validated structural variation. The evaluation must separate the accuracy of the underlying sequence alignment from additional information obtained through contacts or cell attachments. Report breakpoint recall, correspondence ambiguity, chain-map validity, class-survival ranks, localized cycle support, compute and memory costs, and sensitivity to alternative face rules. Compare against ordinary pangenome/structural-variant analysis and an aligned contact-graph baseline. If the 2-complex adds no stable association with measured biological effects after controlling for those baselines, its additional structure has not been justified for that task. These are proposed evaluations; no such cross-genome empirical benchmark is claimed in this manuscript.

This section supplies a conditional algebraic mechanism for comparing changing genomic relational complexes without assuming a global invertible transformation or a solved microscopic field theory. It complements the multiresolution chain maps of Section 16 by making alignment, missing cells, orientation reversal and copy ambiguity explicit. It also sharpens the distinction central to this paper: valid topological correspondences support a comparative analysis of chosen representations, whereas a physical theory of how those representations are generated remains a separate research problem.

18 Probabilistic Genomic Correspondence and the Propagation of Alignment Uncertainty

The partial cellular correspondences of Section 17 make no demand that a genome-wide one-to-one mapping exist. In practice, however, even the choice of the common evidence complex and its two maps may be uncertain. Inversions can have imprecisely resolved breakpoints; segmental

duplications yield competing paralog assignments; multiple pangenome paths may align to the same interval; and contact measurements may disagree about which higher-order cells are justified. Conditioning topological comparison on one favored alignment then conceals an important source of uncertainty. This section develops a finite or countable probabilistic family of *admissible cellular correspondences*, propagates the resulting uncertainty into homology-based quantities, and identifies conditions under which probabilistic summaries are interpretable. It does not assume that a probability over proposed alignments is automatically calibrated by a sequence aligner's mapping score.

Probabilistic sequence alignment is not new: pair-hidden Markov models can provide match-posterior probabilities, as illustrated by ProbCons, and joint Bayesian treatment of alignment and phylogeny has long been pursued in BALi-Phy (Do et al., 2005; Redelings and Suchard, 2005; Suchard and Redelings, 2006). The question here is more specific. We ask how posterior uncertainty over *cellular spans*, *induced maps*, and *face admissibility* changes the claims that can be made about shared genomic topological information. These questions concern the data-informed pathway of Section 14, and require neither a solved group-field action nor a demonstrated condensate.

18.1 Admissible Correspondence Hypotheses and Posterior Weights

Fix two attributed finite cellular complexes $\mathcal{K}_A, \mathcal{K}_B$, a coefficient field $\mathbb{F}$, and a declared construction rule for each type of edge and face. Let $\mathfrak{A}(D)$ be a finite (or countable) collection of alignment-and-correspondence hypotheses compatible with observed data D . Each $a \in \mathfrak{A}(D)$ specifies a common evidence complex $\mathcal{Z}_a$, type-respecting maps of its attributes, and a span of chain maps

$$C_\bullet(\mathcal{K}_A; \mathbb{F}) \xleftarrow{f_{a,\bullet}} C_\bullet(\mathcal{Z}_a; \mathbb{F}) \xrightarrow{g_{a,\bullet}} C_\bullet(\mathcal{K}_B; \mathbb{F}), \quad \partial^A f_a = f_a \partial^{\mathcal{Z}_a}, \quad \partial^B g_a = g_a \partial^{\mathcal{Z}_a}. \quad (62)$$

An admissible hypothesis also records genomic orientation, breakpoint placement, copy assignments, contact or multiway evidence, and which observations are omitted rather than forced into a match. Chain identities are mandatory algebraic checks, but not sufficient biological evidence: maps that invent an unsupported adjacency or equate differently constructed two-cells are excluded by the declared policy. The candidate set may be truncated for computation, in which case omitted posterior mass must be assessed.

A posterior distribution $\pi(a \mid D)$ over these *joint hypotheses* must be defined by a likelihood and prior, or by a validated probabilistic approximation to such a model; assume $\pi_a \geq 0$ and $\sum_a \pi_a = 1$. A collection of independent match probabilities $P(i \leftrightarrow j \mid D)$ generally does *not* define a distribution on globally consistent cellular spans: many individual matches cannot coexist under orientation, copy-number, or chain-closure constraints. In a model with uncertain contact scores ω and uncertain cell-construction policies c , the complete state may instead be $a = (\text{alignment}, \omega, c)$, with conditional law $\pi(a \mid D)$. Uncertainty due to measurement, candidate topology, and the alignment model should be reported separately when possible.

18.2 The Posterior Distribution of Jointly Supported Homology

For each a , let $f_{a,*}, g_{a,*}$ be the induced homology maps and use the jointly supported dimension from Definition 17.1:

$$m_k(a) = \text{rank } f_{a,*} + \text{rank } g_{a,*} - \text{rank} \begin{pmatrix} f_{a,*} \\ g_{a,*} \end{pmatrix} = \dim \frac{H_k(\mathcal{Z}_a; \mathbb{F})}{\ker f_{a,*} + \ker g_{a,*}}. \quad (63)$$

Because the evidence complex and its homology groups may change with a , this is a scalar function on the candidate space, not the dimension of a fixed shared subspace of the two genomes. Its conditional probability mass function and expectation are

$$p_k(r \mid D) = \sum_{a \in \mathfrak{A}(D)} \pi_a \mathbf{1}\{m_k(a) = r\}, \quad \bar{m}_k(D) = \mathbb{E}_\pi[m_k \mid D] = \sum_a \pi_a m_k(a). \quad (64)$$

Credible intervals or probabilities such as $\mathbb{P}_\pi(m_k \geq r \mid D)$ convey alignment-dependent uncertainty more faithfully than a single alignment-specific rank. These are posterior statements *conditional on the candidate set, priors, and observation model*; they do not have an automatic frequentist coverage guarantee.

Proposition 18.1 (Sensitivity to posterior approximation). *Suppose all admissible hypotheses satisfy $0 \leq m_k(a) \leq M_k$ for a finite bound M_k . For any two probability laws $\pi, \tilde{\pi}$ on a common candidate set, let $\text{TV}(\pi, \tilde{\pi}) = \sup_E |\pi(E) - \tilde{\pi}(E)|$. Then*

$$|\mathbb{E}_\pi m_k - \mathbb{E}_{\tilde{\pi}} m_k| \leq M_k \text{TV}(\pi, \tilde{\pi}), \quad |\mathbb{P}_\pi(m_k \geq r) - \mathbb{P}_{\tilde{\pi}}(m_k \geq r)| \leq \text{TV}(\pi, \tilde{\pi}). \quad (65)$$

Proof. For $0 \leq X \leq M_k$, the layer-cake identity gives $\mathbb{E}X = \int_0^{M_k} \mathbb{P}(X > t) dt$. The difference of each pair of event probabilities is at most total variation; integration proves the expectation inequality. The second claim is the defining event bound for total variation. $\square$

If a retained candidate set $\mathfrak{A}_0$ has posterior mass $1 - \delta > 0$ and $\tilde{\pi}$ is the original posterior conditioned on $\mathfrak{A}_0$, then $\text{TV}(\pi, \tilde{\pi}) = \delta$; thus the first error is at most $M_k \delta$. A small posterior tail can still be biologically consequential if it contains a rare rearrangement, even when this scalar bound is small. Bounds on the expectation do not imply that individual matched cycles are identifiable.

18.3 A Finite Ambiguity Example and the Failure of Averaged Maps

Consider $\mathbb{F} = \mathbb{F}_2$, a common evidence complex whose first homology has basis $[u], [v]$, and target first-homology spaces $H_1(\mathcal{K}_A) = H_1(\mathcal{K}_B) = \mathbb{F}_2$. Both hypotheses share the induced map $f_* = (1 \ 0)$, but one maps the first evidence loop into B and the other maps the second:

$$a_1 : \quad f_* = (1 \ 0), \quad g_{1,*} = (1 \ 0); \quad a_2 : \quad f_* = (1 \ 0), \quad g_{2,*} = (0 \ 1). \quad (66)$$

These are realizable by cellular maps between a wedge of two circles and individual circles, with a mapped loop sent either to the target loop or to the basepoint. Formula (63) gives $m_1(a_1) = 1$ and $m_1(a_2) = 0$, even though both hypotheses separately detect one loop in each target. For illustrative weights $\pi(a_1 \mid D) = 0.70$ and $\pi(a_2 \mid D) = 0.30$, the induced law satisfies

$$\mathbb{P}(m_1 = 1 \mid D) = 0.70, \quad \mathbb{E}[m_1 \mid D] = 0.70, \quad \text{Var}(m_1 \mid D) = 0.21. \quad (67)$$

Figure 16 shows that the disagreement concerns *which evidence class is jointly supported*, not whether each target possesses an independent cycle. The probabilities are chosen to illustrate propagation; they were not estimated from genomic data.

Averaging cellular maps is generally an unsuitable substitute for carrying the full hypothesis distribution. When the domains differ, the matrices cannot even be added without arbitrary identifications. When domains and codomains happen to agree, a linear combination of chain maps over $\mathbb{R}$ is again a chain map, but it need not be a cellular map and its induced rank can change discontinuously. For example, two rank-one maps $[u] \mapsto [b]$ and $[u] \mapsto -[b]$ each have rank one, whereas their equally weighted average is zero. This example also demonstrates the danger of mixing orientation conventions: signed representatives may describe the same physical support. In addition, the field $\mathbb{F}_2$ does not support real-valued posterior weighting of chain matrices in the first place. One must not substitute $\text{rank}(\mathbb{E}_\pi[f_{a,*}])$ or $\text{rank}(\mathbb{E}_\pi[g_{a,*}])$ for the well-defined $\mathbb{E}_\pi[m_k(a)]$.

18.4 Joint Uncertainty from Alignment, Contact Evidence, and Filtration

A topology-valued comparison may depend on both the correspondence a and a contact-score realization ω . If $Y_k = Y_k(a, \omega)$ is a bounded scalar summary, then

$$\text{Var}(Y_k \mid D) = \mathbb{E}_{a \mid D}[\text{Var}(Y_k \mid a, D)] + \text{Var}_{a \mid D}(\mathbb{E}[Y_k \mid a, D]). \quad (68)$$

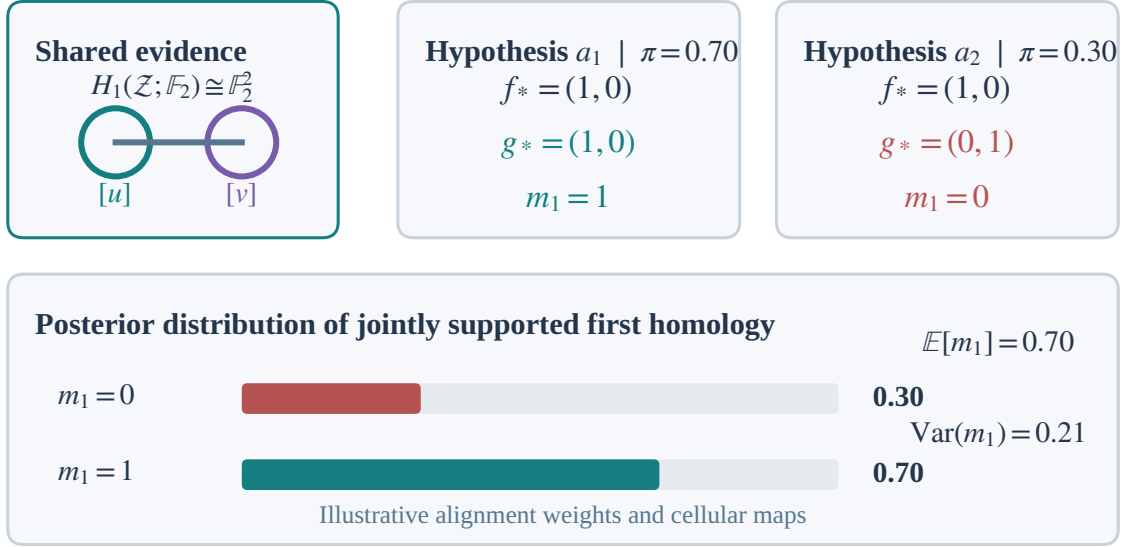


Figure 16: Two possible correspondences for a common evidence complex with two independent first-homology generators. Genome A receives the class $[u]$ in both branches. In hypothesis a_1 , genome B also receives $[u]$, giving a one-dimensional jointly supported quotient. In a_2 , genome B receives $[v]$, giving no jointly supported dimension even though both targets have nontrivial H_1 . The lower panel plots an illustrative posterior distribution over the scalar m_1 (not empirical genomic data).

The first term represents within-correspondence variation under the selected conditional model; the second represents variation between correspondence hypotheses. This is a decomposition within that model, not an automatic causal attribution of uncertainty to alignment versus measurement. Dependence between alignment and contact noise is allowed through the joint posterior. If a different cell-attachment policy is also uncertain, it should be included as a distinct latent modeling choice and subjected to a sensitivity analysis rather than treated as measured randomness.

The persistence-stability estimate of Section 15 applies within a hypothesis only when two filtrations are defined on a *common cellular support* and satisfy the required sup-norm filtration bound. Across different $\mathcal{Z}_a$, changes in boundaries or cell attachments are not controlled by that bound. Consequently, an alignment-averaged barcode is not a canonical persistence module. One may instead compare distributions of scalar persistence summaries, samples of diagrams under a specified metric, or an explicitly defined transport/correspondence between filtered complexes. These options require different assumptions; neither an average of barcode endpoints nor a mean Betti number proves that any particular genomic cycle is conserved.

18.5 Computational Approximation and Calibration Criteria

For a finite candidate set of size R , a direct approach evaluates and stores each admitted span and its $m_k(a)$, with total work on the order of $\sum_{a=1}^R T_{\text{hom}}(a)$ plus the cost of producing the candidate hypotheses. The latter may dominate because full-genome alignments with duplications and rearrangements have combinatorially many alternatives. A practical research implementation may use locus- or block-specific proposals, constrained sampling, sequential Monte Carlo, beam search, or importance weighting; none is guaranteed to explore all biologically plausible modes. Rao–Blackwellization is appropriate only where conditional subproblems can actually be integrated or computed exactly. Cached local homology updates and sparse correspondences may amortize repeated work, but require their own benchmarks.

For independent posterior draws $a^{(1)}, \dots, a^{(R)}$, one can estimate $\bar{m}_k$ by $R^{-1} \sum_r m_k(a^{(r)})$. If

the draws are independent and $0 \leq m_k \leq M_k$, Hoeffding's inequality gives

$$\mathbb{P}(|\hat{m}_k - \bar{m}_k| > \varepsilon \mid D) \leq 2 \exp\left(-\frac{2R\varepsilon^2}{M_k^2}\right). \quad (69)$$

This controls Monte Carlo error *conditional on the specified posterior*, not whether that posterior describes biological truth. It does not apply unchanged to correlated MCMC output or to biased beam-search samples. Such cases require appropriate mixing, weighting, truncation, and Monte Carlo error diagnostics.

Evaluation should begin with spike-in structural variants and simulated alignments of known ground truth, and then move to independently assembled, experimentally supported haplotypes. At minimum, report correspondence recall and ambiguity, calibration of posterior match probabilities, coverage of candidate hypotheses, expected jointly supported ranks and their uncertainty, support of localized cycles, sensitivity to alternate face rules, and runtime and peak memory. Proper scoring rules or reliability plots can assess posterior probabilities when labels are available. Compare the full uncertainty-aware procedure with the maximum-posterior single alignment, a standard pangenome/sequence alignment, and topology computed on common contact graphs without inferred two-cells. A gain in interpretability or held-out biological prediction must survive these comparisons. No genome-pair posterior inference, calibration result, or new software implementation is claimed here.

The substantive conclusion is that *uncertainty over genomic correspondence induces uncertainty over the topological conclusions*. Exact chain identities should be checked for every candidate, but they are not a license to turn uncertain genomic homology into a deterministic correspondence. Carrying candidate spans and their probabilities through the invariant calculation is a concrete route toward comparative genome topology that does not require deriving the microscopic generative action, while preserving a clear path for future coupling to physical dynamics.

19 Joint Probabilistic Inference Across Resolution, Alignment, and Genomic Structure

The previous sections examine contact-score stability on a fixed support (Section 15), cellular information preserved by a declared resolution map (Section 16), structural-variant correspondences (Section 17), and uncertainty over candidate alignments (Section 18). A realistic genome-scale comparison combines all of these choices. The main difficulty is not merely the number of choices, but their dependence: a genomic inversion may change which fine-resolution contacts can be aligned; a coarse partition may conceal a duplication; and a rule for attaching two-cells may depend on uncertain contact scores. This section specifies a joint probability model, proves bounds that remain valid under dependence, and describes which quantities can be compared without silently identifying non-isomorphic cellular objects. It extends the data-informed route independently of the conjectural microscopic group-field action.

19.1 Admissible Joint States and the Status of Resolution Choices

Let $D = (D_A, D_B, D_{\text{meta}})$ denote the paired genome observations, contact measurements, and assay metadata. An analysis state is a tuple

$$z = (r, a, c, \omega, \eta) \in \mathfrak{Z}, \quad (70)$$

where r specifies a pair of binning or multiresolution schemes, a is a partial cellular correspondence hypothesis, c prescribes admissible edge and face constructions, ω comprises latent or uncertain contact scores, and η collects biological and nuisance parameters (such as cell state and assay effects) when modeled. An *admissible* state is one for which both constructed complexes, their

filtrations when invoked, and the typed cellular span have well-defined boundary maps. Thus, for every admitted z , the identities $B_1^A B_2^A = 0$, $B_1^B B_2^B = 0$, and $\partial f_z = f_z \partial$, $\partial g_z = g_z \partial$ are verified or guaranteed by construction. Merely assigning a nonzero probability to an alignment does not make its corresponding cellular map valid.

For finite or suitably measurable $\mathfrak{Z}$, one may specify a normalized joint analysis law

$$q_\lambda(z \mid D) = \lambda(r) q_r(a, c, \omega, \eta \mid D), \quad \sum_r \lambda(r) = 1, \quad (71)$$

where each conditional law q_r is normalized over states allowed at resolution r . This representation does *not* assume that a, c, ω, η are conditionally independent. The weights $\lambda(r)$ deserve special care: genomic bin size and face-attachment policy can be *analyst-selected methodological choices*, not unknown biological random variables. In that case λ is a declared sensitivity-analysis weighting, not a posterior probability that one resolution is biologically true. A genuine Bayesian posterior over a model index would instead require explicit model priors, comparable likelihoods, and treatment of parameter dimensions. The face rule c is likewise a model-choice uncertainty unless a biological stochastic mechanism for cell formation is independently justified.

A shared-homology statistic is calculated only after fixing a complete admissible state. For a selected degree k , let $m_k(z)$ be the quotient dimension defined by Eq. (60) using the span carried by z . If its dimensions depend on r , raw counts are not automatically comparable across resolutions. For a declared common finite cap $M > 0$, report a bounded score $Y_k(z) = \min\{m_k(z), M\}/M \in [0, 1]$, together with the uncapped m_k and its resolution-stratified distribution. Alternatively, choose a reference support with a justified transport map and compare transported classes; neither normalizing by the number of bins nor choosing an arbitrary M establishes biological equivalence. For persistence diagrams the challenge is stronger: diagrams on changing cellular supports need a specified common filtration parameter, convention for essential intervals, and inter-complex comparison rule before a distance is meaningful.

19.2 Joint Propagation and Dependence-Aware Uncertainty Decomposition

For a bounded measurable summary $Y(z) \in [0, M]$, define the joint expectation and a threshold-support probability by

$$\bar{Y}_\lambda = \int_{\mathfrak{Z}} Y(z) q_\lambda(dz \mid D), \quad p_\lambda(t) = q_\lambda\{z : Y(z) \geq t \mid D\}. \quad (72)$$

The distribution of Y is a pushforward of the *joint* law: it is not obtained by averaging a collection of chain matrices, by separately averaging persistence bars, or by multiplying unrelated marginal distributions. For multiple homology degrees or localized classes, a vector of jointly computed summaries should be carried through the same sampling hypotheses to preserve their posterior dependence.

Proposition 19.1 (Nested variance decomposition with dependent structural hypotheses). *Suppose Y has finite second moment under q_λ and the stated conditional expectations exist. Then, with R the resolution index and A the correspondence index,*

$$\text{Var}(Y \mid D) = \mathbb{E}_{R,A \mid D}[\text{Var}(Y \mid R, A, D)] \quad (73)$$

$$\text{notag} + \mathbb{E}_{R \mid D}[\text{Var}_{A \mid R, D}(\mathbb{E}[Y \mid R, A, D])] \quad (74)$$

$$\text{notag} + \text{Var}_{R \mid D}(\mathbb{E}[Y \mid R, D]). \quad (75)$$

The first conditional term includes uncertainty in contact scores, cell rules, and other remaining latent variables. The decomposition permits arbitrary dependence between correspondence and resolution through $q_\lambda(A \mid R, D)$.

Proof. Apply the law of total variance first conditioning on (R, A) , then apply it again to the random variable $\mathbb{E}[Y \mid R, A, D]$ conditioning on R . No independence assumption is needed. $\square$

These are *nested conditional components*, not uniquely identified causal contributions. Their numerical values change with the order of conditioning, and the third term reflects the selected resolution mixture λ when resolution is an analysis decision. Reporting a single total variance without the corresponding conditionals would conceal that distinction.

Proposition 19.2 (Robustness of a joint topological summary). *Let $q, \tilde{q}$ be probability laws on the same admissible state space, and suppose $0 \leq Y(z) \leq M$. Then*

$$\left| \mathbb{E}_q Y - \mathbb{E}_{\tilde{q}} Y \right| \leq M \text{TV}(q, \tilde{q}), \quad |q(Y \geq t) - \tilde{q}(Y \geq t)| \leq \text{TV}(q, \tilde{q}). \quad (76)$$

If q is restricted and renormalized to a set of total mass $1 - \delta > 0$, the expectation error is at most $M\delta$.

Proof. The layer-cake representation for nonnegative bounded functions proves the first inequality; the second is the defining event bound for total variation. The total variation between q and its conditional restriction is exactly δ . $\square$

This result extends Proposition 18.1 to joint structural hypotheses, but it is not a stability theorem for unaligned persistence diagrams. The scalar bound requires a globally bounded statistic on a *common admissible hypothesis space*; it provides no guarantee that a truncated candidate space includes the correct biological rearrangement, nor any guarantee about model misspecification.

19.3 A Finite Example in Which Independent Marginals Give the Wrong Conclusion

Consider two resolution strategies $R \in \{r_0, r_1\}$ and two admissible correspondence hypotheses $A \in \{a_0, a_1\}$. Let their declared joint law be

$$\begin{array}{c|cc} & a_0 & a_1 \\ \hline r_0 & 0.45 & 0.05 \\ r_1 & 0.05 & 0.45 \end{array}, \quad \mathbb{P}(R = r_i) = \mathbb{P}(A = a_i) = \frac{1}{2}. \quad (77)$$

For the complete states indexed by (r_i, a_j) , suppose the selected cellular spans have $W_1 = \mathbb{F}_2^2$, $A_1 = B_1 = \mathbb{F}_2$, $f_* = (1 \ 0)$, and $g_* = (1 \ 0)$ when $i = j$ but $g_* = (0 \ 1)$ when $i \neq j$. These are legitimate homology maps realized by maps between wedges of circles and single circles, as in Section 18. Their shared-rank statistic is therefore $m_1(r_i, a_j) = \mathbf{1}\{i = j\}$. Under the joint law,

$$\mathbb{E}[m_1] = 0.90, \quad \text{Var}(m_1) = 0.09. \quad (78)$$

Replacing the joint law by the product of its *identical marginals* assigns probability 0.25 to every pair, yielding $\mathbb{E}_{q_R q_A}[m_1] = 0.50$ and variance 0.25. Thus matching marginal resolution and alignment distributions do not determine a valid joint topological expectation. The dependence between these choices changes the answer by 0.40 in this intentionally synthetic example (Figure 17). It is not a genomic measurement; its purpose is to demonstrate a logically necessary modeling distinction that matters whenever valid correspondences depend on the representation scale.

A Dependent hypotheses at two resolutions

		Alignment		
		a_0	a_1	
r_0		<div><div>.45</div><div>$m_1 = 1$</div></div>	<div><div>.05</div><div>$m_1 = 0$</div></div>	Joint expectation $\mathbb{E}[m_1] = 0.90$
r_1		<div><div>.05</div><div>$m_1 = 0$</div></div>	<div><div>.45</div><div>$m_1 = 1$</div></div>	Independent marginals $\mathbb{E}_{q_R q_A}[m_1] = 0.50$

B Verify each cellular hypothesis before aggregation

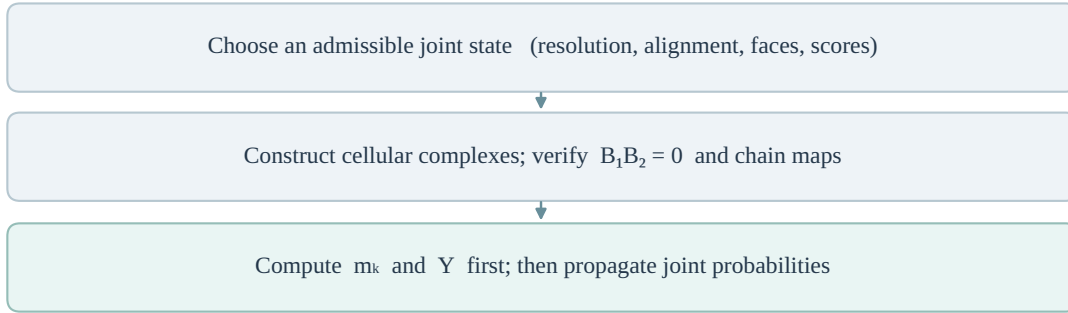


Figure 17: Dependence-aware joint topological inference in a finite illustrative example. The two-by-two table displays $q(R, A | D)$, and the upper-right annotations show the shared-homology rank m_1 : matching pairs preserve one class; off-diagonal pairs preserve none. Independent marginal multiplication would replace the displayed 0.45/0.05 probabilities by 0.25 everywhere, changing the expected common dimension from 0.90 to 0.50. The right-hand diagram distinguishes admissibility checks and within-hypothesis homological computation from the final uncertainty pushforward. Neither the probabilities nor the complexes are inferred from real genomes.

19.4 Propagation to Persistent Features and Cross-Resolution Comparisons

For every joint state z , the contact realization ω and face rule c determine a filtration $\mathcal{K}_z^\lambda$ and its persistence module $H_k(\mathcal{K}_z^\lambda)$ when a valid filtration exists. The posterior object is consequently a distribution over persistence modules or diagrams, not automatically a single module. Within a fixed support and face rule, Section 15 supplies a bottleneck stability bound for small score errors; across different supports, scales, or alignment hypotheses, no such bound follows without additional interleavings, comparison maps, or common refinements. In particular, choosing a minimum-distance matching of barcode endpoints does not supply a biological locus-to-locus correspondence.

A defensible practical summary is a distribution of prespecified scalar quantities evaluated consistently within each state, such as the number of bars above a normalized persistence threshold, distance to an independently selected reference after defining compatible scales, or the proportion of reference-anchored classes carried by verified chain maps. A topological comparison should show how those summaries change when one conditions on a resolution r , varies the face rule c , or excludes poorly supported alignment modes. For genome pairs with nonmatching contact assays, depth, cell composition, or batch processing, the nuisance model η must be calibrated or results reported as conditional on those uncontrolled factors. Comparative contact-map benchmarking

has shown that different methods prioritize different kinds of map changes, reinforcing the need for prespecified endpoints and held-out checks (Gjoni et al., 2025; Fletez-Brant et al., 2024).

19.5 Scalable Inference and an Empirical Validation Protocol

Exact summation over resolution schemes, correspondence alternatives, face rules, and contact-score realizations generally has a product-sized state space even when the individual cellular complexes are sparse. A prospective genome-scale implementation should separate *candidate generation* from *conditional topological evaluation*. At each declared resolution, propose biologically admissible alignment and structural hypotheses, verify their typed boundary and chain-map identities, calculate or approximate only the requested invariant on sparse cellular representations, and retain the candidate probability or importance weight with that invariant. Conditional elimination of contact variables or local cellular calculations may reduce work, but should be justified for the specific model. Approximation by pruning, importance sampling, MCMC, or amortized proposals creates a second layer of computational error; posterior-mass bounds, effective-sample-size diagnostics, and repeated-chain checks should accompany it. Correlated samples cannot be assigned independent-sample concentration bounds without additional assumptions.

For an empirical study, the necessary validation conditions are distinct: (i) controlled or simulated ground-truth structural variants to validate alignment candidates and copy assignments; (ii) experimentally replicated Hi-C/Micro-C data to evaluate contact-score and normalization uncertainty; (iii) held-out regions or cell conditions to test predictive relevance of localized topological differences; and (iv) independently assembled haplotypes or verified genome comparisons to assess between-genome transfer. An analysis should compare a full joint model to independently factorized uncertainty, the maximum-posterior correspondence at a single resolution, contact-graph-only topology, and established non-topological contact-map comparisons. Report calibration for observable outcomes, sensitivity to $\lambda(r)$, the mass of discarded hypotheses, runtime, and peak memory. These proposed tests have not been executed on biological datasets in the present work.

The methodological result is thus conditional but useful: a genome-scale topological study can carry the uncertainties of resolution, structural correspondence, and contact-derived cellular construction jointly, without assuming that the observed complex emerged from a specified microscopic field action. The exact finite identities quantify where seemingly innocuous marginal averaging or deterministic correspondence can fail. Establishing which richer structural variables are biologically identifiable and predictive remains an empirical task.

20 Identifiability and Information Limits in Generative Genomic Reconstruction

The preceding analysis has made explicit several sources of uncertainty: contact evidence, cellular face attachment, changes of genomic resolution, and uncertain correspondences between distinct genome assemblies. A deeper obstacle remains even when all those numerical uncertainties are reduced. A proposed generative system may admit *distinct internal mechanisms that generate exactly the same distribution of measured observables*. In that case no improvement in sequence sample size, numerical precision, or optimization method can recover which mechanism was responsible from those observables alone. Conversely, uncertainty about an individual cellular history must not be mistaken for nonidentifiability of a statistical parameter, and gauge-related descriptions must not be treated as physically distinct histories. This section develops these distinctions for the finite contact model, establishes an exact nonidentifiable parameter family, and specifies which additional measurements could distinguish its members. The constructions are mathematical results for the stated model, not claims that a real genome is generated by this particular equilibrium process.

20.1 Observation Maps, Physical Equivalence, and Statistical Identifiability

Let $z = (\mathcal{K}, \mathcal{H}, U, \Phi, c, \vartheta)$ denote a latent generative description, including a cellular support, a possible history, internal fields, contacts, and parameters. A measurement protocol d produces observations Y_d through a likelihood $p_d(y | z)$. The experimentally accessible law is obtained by marginalizing the unobserved structure,

$$p_{\vartheta,d}(y) = \int p_d(y | z) \mu_{\vartheta}(dz), \quad (79)$$

where the integral includes finite sums when appropriate. The family μ_{ϑ} and the observation kernel must be specified before claims of identifiability can be evaluated. For a fixed measurement protocol define the observational equivalence relation

$$\vartheta \sim_d \vartheta' \iff p_{\vartheta,d}(y) = p_{\vartheta',d}(y) \text{ for all admissible } y. \quad (80)$$

A parameter functional $T(\vartheta)$ is identifiable under d if $\vartheta \sim_d \vartheta'$ implies $T(\vartheta) = T(\vartheta')$; full parameter identifiability additionally requires that the equivalence classes be singletons after legitimate symmetries and irrelevant reparametrizations have been removed. This distinction parallels established identifiability analyses of latent-state models (Allman et al., 2009; Cole, 2019).

There are at least four different obstructions in the present research program. First, a local gauge transformation changes (h, U) while leaving the physical composites q unchanged; the quotient in Proposition 5.2 removes this *representational* redundancy. Second, the four-state sequence can be observed while latent contacts remain unmeasured, producing *statistical* equivalence between genuinely different coupling parameters. Third, different cellular fillings or histories may share the same terminal boundary and observation law; this is a *structural or historical* inverse problem beyond the gauge quotient. Fourth, topological summaries such as Betti numbers or persistence diagrams further discard information and need not distinguish different constructed complexes, even when the complexes are individually well defined. One cannot infer a unique physical spacetime from a matching collection of topological invariants.

20.2 An Exact Nonidentifiable Family in the Contact-Coupled Model

Write $S_{\text{ct}}(x) = \sum_{(i,j) \in E_{\star}} \delta_{ij}(x)$ and $S_{\text{seq}}(x) = \sum_i \delta_{i,i+1}(x)$ for a fixed, nonempty candidate contact set $E_{\star}$. Equations (17)–(19) imply

$$p_{K,\nu}(x) = \frac{\exp[\beta J S_{\text{seq}}(x) + \alpha(K, \nu) S_{\text{ct}}(x)]}{Z_{\text{seq}}(\alpha, J)}, \quad \alpha(K, \nu) = \log \frac{1 + e^{\beta(K-\nu)}}{1 + e^{-\beta\nu}}. \quad (81)$$

The sequence-independent contact factor has been absorbed into Z_{seq} , so $\alpha = \beta \Delta_{\text{contact}}$ is the *only* combination of K and ν retained by the marginal sequence law when $\beta, J, E_{\star}$ are held fixed.

Proposition 20.1 (An exact sequence-observation equivalence class). *Fix $\beta > 0$, J , and $E_{\star}$. If $\alpha(K, \nu) = \alpha(K', \nu')$, then $p_{K,\nu}(x) = p_{K',\nu'}(x)$ for every word x and every system size with the declared candidate-edge construction. For a prescribed $\alpha > 0$, the positive-reward parameterizations on this level set satisfy*

$$K(\nu; \alpha) = \nu + \beta^{-1} \log[e^{\alpha}(1 + e^{-\beta\nu}) - 1]. \quad (82)$$

If S_{ct} is nonconstant over the possible words, the scalar α itself is identifiable from the entire exact sequence distribution at fixed $\beta, J, E_{\star}$, but the two separate parameters (K, ν) are not.

Proof. For $\delta \in \{0, 1\}$, write $1 + e^{\beta(K\delta - \nu)} = A_0 e^{\alpha\delta}$ with $A_0 = 1 + e^{-\beta\nu}$, where α is Eq. (81). Multiplication over all $|E_{\star}|$ candidate contacts yields the factor $A_0^{|E_{\star}|}$, independent of x ; it cancels when normalizing $p(x)$. Solving the defining equation for $e^{\beta(K-\nu)}$ gives Eq. (82). If two values $\alpha \neq \alpha'$ produced the same normalized law, their probability ratio would be constant across x , forcing $(\alpha - \alpha')S_{\text{ct}}(x)$ to be constant, contrary to the stipulated nonconstancy of S_{ct} . $\square$

The loss of a parameter direction also appears in local information geometry. Let $\theta = (K, \nu)$, let $\sigma(t) = 1/(1 + e^{-t})$, and define $p_1 = \sigma(\beta(K - \nu))$ and $p_0 = \sigma(-\beta\nu)$. Differentiation gives

$$\nabla_{\theta}\alpha = \beta \begin{pmatrix} p_1 \\ p_0 - p_1 \end{pmatrix}, \quad \mathcal{I}_x(\theta) = \text{Var}_{p_{\theta}(x)}(S_{\text{ct}}) (\nabla_{\theta}\alpha)(\nabla_{\theta}\alpha)^{\top}. \quad (83)$$

Thus the Fisher-information matrix of a single exact sequence observation has rank at most one for the two-dimensional contact parameter θ ; for independent observations it remains rank at most one after summation. If S_{ct} varies with the word and $K > 0$, the rank is exactly one. This is a *structural* singularity, not a deficiency that more precise numerical optimization will fix. If both J and β are also unknown, additional confounding must be analyzed rather than assuming that they are known or separately calibrated.

20.3 Conditional Contact Observations and an Explicit Finite Example

Contact measurements change the observation law. By Eq. (16), for candidate contacts whose terminal states disagree or agree, respectively,

$$p_0 = \mathbb{P}(c_{ij} = 1 \mid x_i \neq x_j) = \sigma(-\beta\nu), \quad p_1 = \mathbb{P}(c_{ij} = 1 \mid x_i = x_j) = \sigma(\beta(K - \nu)). \quad (84)$$

Proposition 20.2 (Conditional identification with observed binary contacts). *Assume $\beta > 0$ is known, conditional contact probabilities $p_0, p_1 \in (0, 1)$ are available under a calibrated observation mechanism, and both equality classes occur with nonzero probability among the measured candidate edges. Within the specified equilibrium model, (K, ν) is then identifiable from the pair (p_0, p_1) , with*

$$\nu = -\beta^{-1} \text{logit}(p_0), \quad K = \beta^{-1} [\text{logit}(p_1) - \text{logit}(p_0)]. \quad (85)$$

If two distinct parameter pairs have the same sequence marginal, their joint distributions of sequence and the complete calibrated contact vector nevertheless differ whenever the conditional contact probabilities differ for a candidate-edge equality class with positive sequence probability.

Proof. Invert the two logistic expressions in Eq. (84). For the final assertion, the sequence marginal cancels in the chain rule for relative entropy,

$$D_{\text{KL}}(p_{\theta}(x, c) \parallel p_{\theta'}(x, c)) = \mathbb{E}_{p_{\theta}(x)} \sum_{e \in E_{\star}} D_{\text{KL}}(\text{Bern}(p_{\delta_e}) \parallel \text{Bern}(p'_{\delta_e})). \quad (86)$$

Every term is nonnegative, and a term is strictly positive when its two Bernoulli parameters differ; the stipulated event has positive probability. Equality of the sequence laws was used in the displayed identity. The general chain rule would otherwise include the divergence of the sequence marginals. $\square$

Figure 18 visualizes this obstruction with $\beta = 1$, $J = 0.35$, and two parameter sets, $(K, \nu) = (1.2, 0.4)$ and $(K', \nu') \simeq (1.708868, 1.35)$. Both give $\alpha \simeq 0.658085$, so their entire sequence distributions coincide for the same candidate contact graph, not merely one or two summary statistics. Yet they yield $p_0 \simeq (0.4013, 0.2059)$ and $p_1 \simeq (0.6900, 0.5888)$ for the two parameterizations. Complete enumeration with four, five, and six terminal positions confirms normalized sequence-distribution agreement to floating-point precision. The estimated values in this example are exactly calculated model probabilities, not measured genomic contact rates.

The proposition concerns *parameter recoverability under a specified joint distribution*. It does not recover the realized contact pattern of a past cell from a terminal word. Even if K and ν are known, each edge still has a nondegenerate conditional Bernoulli posterior in Eq. (16). Biological Hi-C and related assays also report noisy, protocol-dependent measurements rather than the exact binary c used above. Hi-C, GAM, and SPRITE can differ in their sensitivity to chromatin structures and cell-to-cell variability (Fiorillo et al., 2021); therefore Eq. (85) cannot simply be applied to an uncalibrated contact-frequency table.

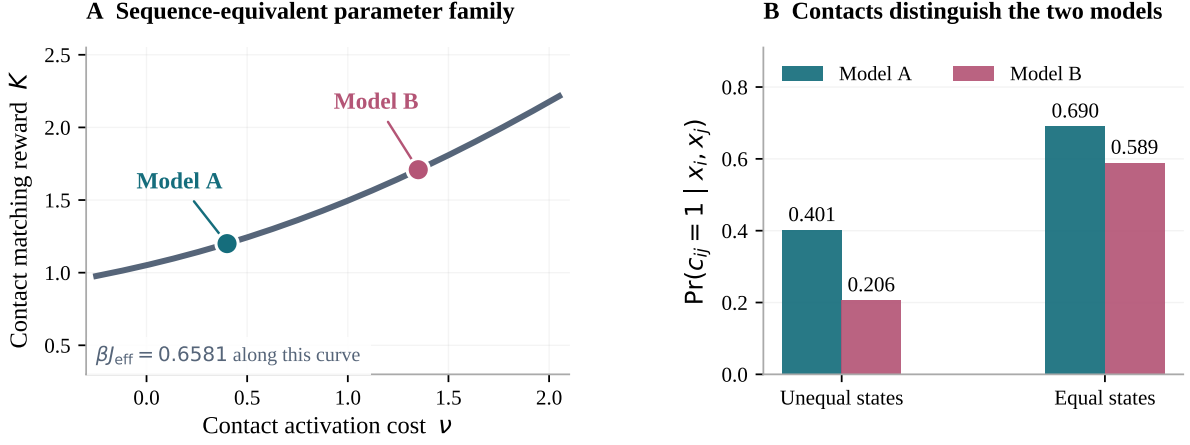


Figure 18: An exact sequence-only identifiability obstruction. **A:** a continuous level set of equal effective sequence coupling $\alpha = \beta \Delta_{\text{contact}}$, with two distinct microscopic (K, ν) specifications. **B:** the same models predict different conditional binary-contact rates for unequal and equal sequence states, respectively. The model illustrates how adding a calibrated contact observable can distinguish generative mechanisms that sequence data alone cannot distinguish; it is not a fitted result on biological genomes.

20.4 Information Bounds and the Choice of Additional Measurements

The practical consequence of observational equivalence can be phrased without assuming a particular parameter estimator. Suppose a finite hypothesis variable $\Theta \in \{\theta_A, \theta_B\}$ assigns positive prior probability to two different mechanisms for which $p_{\theta_A}(x) = p_{\theta_B}(x)$. Any number n of conditionally independent sequence samples drawn under either model then satisfies

$$I(\Theta; X^{(1)}, \dots, X^{(n)}) = 0, \quad (87)$$

with the model-index prior left unchanged by sequence-only observations. A statistic deterministically computed from those sequences, such as a sequence-constructed fixed-complex Betti vector, cannot recover the lost distinction by the data-processing inequality. The statement does not apply to a new assay that supplies genuinely additional observations, or to a complex constructed using further independent physical evidence.

For a candidate measurement design d generating an observation M_d , an information-focused objective is

$$\text{EIG}(d | D) = I(\Theta; M_d | D, d) = \mathbb{E}_{M_d | D, d} [D_{\text{KL}}(p(\Theta | D, M_d, d) \| p(\Theta | D))]. \quad (88)$$

In a design with different assay expenses, one can maximize information gain subject to cost and accessibility constraints. This is an instance of Bayesian experimental design, already studied in spatial genomics (Jones et al., 2024). For the finite binary-contact example, observing calibrated contacts in both equality strata is informative about separate (K, ν) values; additional terminal sequences drawn from the same marginal are not. For physical genomic reconstruction, candidate complementary observables might include condition-matched contact data, multiway spatial contacts, locus-specific imaging, controlled structural perturbations, or temporally resolved data. Their actual information gain must be estimated under an assay-aware likelihood, not asserted from modality names alone.

An intervention adds a different kind of test. A perturbation of a molecular mechanism requires its own causal or kinetic response law; merely changing a conditioned symbol in a fixed equilibrium distribution is not an intervention. Distinct generative theories may make identical observational predictions but different intervention predictions, provided the interventions and their effects on the underlying model are defined. Similarly, a cellular-history theory must specify which

measurements can distinguish different interior histories sharing an identical boundary readout; without that specification the inverse history claim is underdetermined.

20.5 Topological Reconstruction and Identifiable Scientific Targets

For data-informed two-complexes, the inverse problem should be stated at the resolution of the scientific target. Let $\mathcal{T}$ be a complex-construction rule, $\mathcal{O}$ an assay model, and $\Phi_k(\mathcal{K})$ a chosen topological descriptor. One may aim to recover $\Phi_k(\mathcal{K})$ without recovering $\mathcal{K}$ up to isomorphism, or to recover a conserved subcomplex without identifying the full generative history. A property $F(z)$ is *observation-identifiable* only if it is constant on every latent equivalence class under the declared observational law; stability of a property under score perturbations (Section 15) is not the same property as identifiability. Likewise, an exact cellular chain map (Section 16) does not imply that the underlying fine-resolution structure can be uniquely inverted.

This distinction motivates three different reconstruction goals. A *descriptive* target is a stable annotated topological statistic computed from a declared input construction; it does not claim to be an intrinsic unique invariant of an unobserved physical genome structure. An *inferential* target is a posterior distribution over complexes or correspondences conditional on explicit assay models, reporting nonidentifiability when multiple admissible structures fit the data. A *mechanistic* target is a predictive causal or dynamical law that is constrained by independent measurements, perturbations, and model comparisons. For example, homology signatures cannot by themselves distinguish two distinct cellulations with the same homology groups; there is no guarantee that increasingly elaborate cohomological features eventually determine a unique causal history. Information must be supplied by additional assumptions or observations rather than manufactured by algebraic postprocessing.

A serious genomic implementation should therefore identify its target estimand before choosing its topological descriptor, distinguish positional and biological correspondences from topological equivalence, and report at least (i) competing latent explanations, (ii) likelihood or posterior calibration under simulated known-ground-truth controls, (iii) sensitivity to assay noise and alternative face rules, (iv) parameter-recovery tests with and without complementary data, and (v) counterexamples establishing what remains nonidentifiable. The most useful near-term contribution may be to identify a stable, experimentally predictive *quotient-level* property rather than an imagined unique microscopic genomic spacetime. A later claim of an end-to-end condensate mechanism would require the stronger task of showing that its observationally testable predictions cannot be reduced to equally successful simpler descriptions.

21 Distinguishing Generative Histories Through Longitudinal and Intervention-Based Genomic Observations

The preceding identifiability analysis concerns equilibrium sequence and contact measurements. A generative account of genomic structure also requires the more difficult question of whether distinct *histories* or transition mechanisms can be distinguished by observations made at several times, and whether a controlled perturbation provides information unavailable under passive observation. We address this problem with a finite-state controlled stochastic model. It is an effective kinetic extension of the finite relational framework, not a derivation of its dynamics from the proposed group-field action. In particular, an arbitrary sequence of observed contact maps does not determine a unique history of cellular complexes: the observation operator, sampling design, and admissible transitions must be specified.

21.1 Controlled Dynamics on Relational Configurations

Let $\mathbf{X}$ be a finite set of *attributed cellular configurations*. A state $x \in \mathbf{X}$ may include a two-complex $\mathcal{K}_x$, a boundary word w_x , internal contact or chemical attributes, and a correspondence to the preceding cellular configuration. The finite-state assumption is an explicit approximation; it permits exact calculations but does not identify the continuum or GFT state space with a finite Markov system. Let u index an externally specified environmental or experimental condition. For each u , let $Q_\theta^{(u)} = (Q_{xy}^{(u)})$ be a continuous-time Markov generator on $\mathbf{X}$, satisfying

$$Q_{xy}^{(u)} \geq 0 \quad (x \neq y), \quad Q_{xx}^{(u)} = - \sum_{y \neq x} Q_{xy}^{(u)}, \quad P_{\theta,t}^{(u)} = \exp(tQ_\theta^{(u)}). \quad (89)$$

The generator represents the currently admissible transitions and their rates. In a future genomic implementation these may include contact formation and dissolution, switching of regulatory states, or appropriately represented sequence/structural transformations; one must not assume they all occur on the same biological timescale. When the cell structure changes, any comparison of attributes between $\mathcal{K}_x$ and $\mathcal{K}_y$ must use an explicitly declared cellular correspondence, as in Section 17. If the control changes between observation times, its generator must also change; it is not mathematically equivalent to conditioning on a covariate after observation.

Observed data lie in a different space $\mathbf{Y}$, such as a sequence readout, noisy contact measurements, localized chromatin trajectories, or topological descriptors of an inferred complex. Let $E_\theta(y | x, u)$ be the emission distribution. For observation times $0 = t_0 < t_1 < \dots < t_m$, controls held constant at u_j during $(t_{j-1}, t_j]$, and observations conditionally independent given the latent states, the finite-state likelihood takes the form

$$\begin{aligned} p_\theta(y_{0:m} | \text{do}(u_{1:m})) \\ = \sum_{x_0, \dots, x_m} p_\theta(x_0) E_\theta(y_0 | x_0) \prod_{j=1}^m \left[\exp\left((t_j - t_{j-1}) Q_\theta^{(u_j)}\right) \right]_{x_{j-1} x_j} E_\theta(y_j | x_j, u_j). \end{aligned} \quad (90)$$

An observation taken after a destructive assay is not a later observation of the *same* cell. In that setting the likelihood requires an experimental sampling model, potentially using lineage relations or replicated population cohorts. The simple trajectory likelihood above applies only when individual trajectories or their justified linked representations are available.

Definition 21.1 (Controlled observational equivalence). *For a declared family of admissible experimental designs $\mathfrak{U}$, two parameterized generative models M and M' are observationally equivalent under $\mathfrak{U}$ if they induce the same distribution of every observable dataset under every design in $\mathfrak{U}$. A design comprises the initial-state preparation, control schedule, sampling times, and observation channels. Equivalence under a restricted family need not imply equivalence under an enlarged family with a newly available intervention or measurement.*

This definition avoids treating a reconstructed trajectory, a sequence, and a measurement of topology as interchangeable kinds of data. It also distinguishes *predictive separation of specified hypotheses* from unique recovery of every microscopic parameter.

21.2 Stationary Equivalence Does Not Imply Dynamical Equivalence

Consider the two-state symmetric continuous-time chain with an observable binary readout $q \in \{0, 1\}$ and flip rate $r > 0$:

$$Q_r = \begin{pmatrix} -r & r \\ r & -r \end{pmatrix}, \quad \pi_r = (1/2, 1/2), \quad \Pr(q_t = q_0 | q_0) = \frac{1 + e^{-2rt}}{2}. \quad (91)$$

All positive values of r have the same stationary distribution, whereas their longitudinal relaxation functions differ. Thus, even *noise-free repeated equilibrium snapshots* identifying π_r exactly cannot

recover r ; a time-linked transition observation at a known elapsed time can distinguish specified rates. This is an elementary identifiability result, not a new property of genome chemistry. The corresponding temporal information also depends on measurement error, observation frequency, and whether the sampling protocol preserves the same dynamical population.

A further obstruction arises when the observed sequence is a projection of a larger relational state. Even if X_t is Markovian, its readout $Y_t = b(X_t)$ generally is not. The necessary and sufficient condition for its Markovianity *for every initial distribution* is the familiar strong lumpability condition for the partition into fibers of b (Ball and Yeo, 1993). The following formulation makes its significance explicit for genomic readouts.

Proposition 21.2 (Strong lumpability of an exact cellular readout). *Let Q be a finite-state continuous-time generator, and let $b : \mathbf{X} \rightarrow \mathbf{Y}$ be a deterministic, surjective readout. Its projected process $b(X_t)$ is a time-homogeneous Markov chain for every initial law on $\mathbf{X}$, with a single common projected generator, if and only if for every $a \in \mathbf{Y}$, every $x, x' \in b^{-1}(a)$, and every $c \neq a$,*

$$\sum_{z \in b^{-1}(c)} Q_{xz} = \sum_{z \in b^{-1}(c)} Q_{x'z}. \quad (92)$$

When this holds, the common value defines the off-diagonal entry $\bar{Q}_{ac}$, and the diagonal is determined by zero row sums.

Proof. Under the stated equality the instantaneous rate of entering any other observed class c depends only on $a = b(x)$, not on which hidden state x is occupied. Summing the forward equations over every readout fiber therefore yields a closed forward equation with generator $\bar{Q}$ for every initial law. Conversely, if the readout admits the same Markov generator for every initial law, start the underlying chain at either x or x' in the same fiber and compare the first-order small-time probabilities of the observed transition to class c ; their derivatives at zero are exactly the two sums in (92) and hence must agree. $\square$

Failure of this criterion does not rule out weak lumpability for special initial laws, nor does it make a full hidden-state model incorrect. Rather, it warns that imposing a first-order Markov model directly on sequence symbols may discard memory induced by unobserved relational states. In particular, a topology-valued descriptor estimated at each time point need not inherit Markovian dynamics from the latent complex process.

21.3 An Exact Intervention That Separates Hidden Relational Mechanisms

We now construct two controlled models having identical observed sequence trajectories under baseline conditions and identical *single-time stationary readouts* even under intervention, but distinct intervention-dependent longitudinal observations. Let $x = (h, q)$, where $h \in \{0, 1\}$ is a static unobserved relational class (potentially representing different attributed complexes) and $q \in \{0, 1\}$ is the observable binary readout. The two observed values can be embedded as two distinguishable states of the four-letter readout; this minimal example is not a four-base chemical model. Its symbolic flips are not assertions about the physical rates of nucleotide substitution; a biological application would distinguish mutable sequence states from more rapidly varying chromatin or regulatory readouts. Initially, $\Pr(h = 0) = \Pr(h = 1) = 1/2$ and $q_0 = 0$, independently of h . During the experiment h does not switch, whereas q changes symmetrically at rate $r^{(u)}(h)$. In both models A and B the baseline rates are

$$r_A^{(0)}(0) = r_A^{(0)}(1) = r_B^{(0)}(0) = r_B^{(0)}(1) = 1. \quad (93)$$

Under the designed intervention $u = 1$, set

$$(r_A^{(1)}(0), r_A^{(1)}(1)) = (0.5, 3.5), \quad (r_B^{(1)}(0), r_B^{(1)}(1)) = (2, 2). \quad (94)$$

The intervention is an abstract alteration of the state-conditioned rates; demonstrating a biological counterpart would require independently verified perturbation targets and kinetics.

Proposition 21.3 (Baseline equivalence and intervention separation). *For the two models (93)–(94): (i) all finite-dimensional distributions of the observed baseline process q_t agree; (ii) the intervention-conditional stationary readout is uniform in both models, but the transient retention probabilities satisfy, for $t > 0$,*

$$\begin{aligned} R_A(t) &:= \Pr_A(q_t = q_0 \mid q_0 = 0, u = 1) = \frac{1}{2} + \frac{1}{4}(e^{-t} + e^{-7t}), \\ R_B(t) &:= \Pr_B(q_t = q_0 \mid q_0 = 0, u = 1) = \frac{1}{2} + \frac{1}{2}e^{-4t}, \end{aligned} \quad (95)$$

and $R_A(t) > R_B(t)$. Moreover the projected intervention process in model A is not a time-homogeneous two-state Markov chain for the stipulated hidden-state mixture, whereas the projection of model B is.

Proof. Under $u = 0$, the flip rate of q is one independently of h in both models, proving equality of the entire observed baseline path law. For either hidden class, the symmetric generator (91) has the uniform stationary distribution; mixing over h preserves it. Conditioned on h and $q_0 = 0$, the probability of retaining q_0 is $(1 + e^{-2r(h)t})/2$. Averaging over the two hidden classes gives (95). Strict convexity of the exponential shows $\frac{1}{2}(e^{-t} + e^{-7t}) > e^{-4t}$ for $t > 0$. To establish non-Markovianity, write $C_A(t) = 2R_A(t) - 1 = (e^{-t} + e^{-7t})/2$. Then

$$C_A(2t) - C_A(t)^2 = \frac{1}{4}(e^{-t} - e^{-7t})^2 > 0, \quad (96)$$

whereas a homogeneous symmetric two-state Markov chain has a single-exponential correlation and satisfies the semigroup identity $C(2t) = C(t)^2$. Equivalently, model A violates (92) because its exit rates from the two hidden states with the same readout differ, whereas model B satisfies the condition. $\square$

At $t = 0.6$, the exact prediction is $R_A(0.6) \simeq 0.64095$ versus $R_B(0.6) \simeq 0.54536$ (Figure 19). Direct evaluation of the four-state matrix exponential verifies both expressions to numerical precision. The difference results from *hidden kinetic heterogeneity*, not a demonstrated phase transition or topology-changing genomic event. As in Section 20, the conclusion concerns the distinguishability of the two specified models under a particular intervention; it does not uniquely reconstruct which two-complexes or microscopic fields generated a measured cell trajectory.

21.4 Information Limits and Experimental Design for Time-Linked Data

For a fixed design ξ specifying controls, observation times, initial preparations, and assay channels, denote the induced observed path laws under two hypotheses by $\mathbf{P}_A^\xi$ and $\mathbf{P}_B^\xi$. Distinguishability can be quantified by, for example, their total variation distance or the Kullback–Leibler divergence $D_{\text{KL}}(\mathbf{P}_A^\xi \parallel \mathbf{P}_B^\xi)$. These metrics depend on the full observation process, not only the latent generator. If the laws are identical under all experimentally permissible designs, no estimator operating on those observations can distinguish the hypotheses; the problem is structurally unidentifiable within that observation and intervention family.

For the experiment reporting only whether $q_t = q_0$, the outcome is Bernoulli with parameter $R_M(t)$ under model M . The observed divergence is therefore

$$\begin{aligned} D_{\text{KL}}^\xi(A \parallel B) &= R_A \log \frac{R_A}{R_B} \\ &\quad + (1 - R_A) \log \frac{1 - R_A}{1 - R_B}. \end{aligned} \quad (97)$$

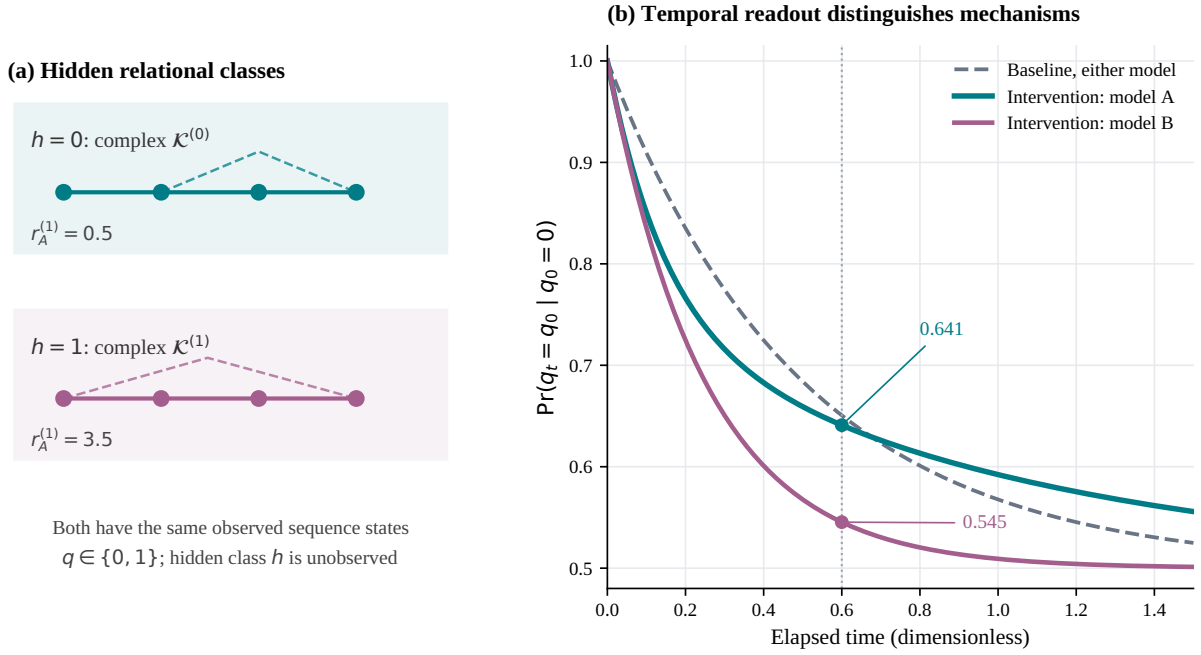


Figure 19: An illustrative intervention-based distinction between two sequence-equivalent baseline mechanisms. Panel (a) depicts two hypothetical hidden relational classes carrying the same binary readout. Under baseline control both models have observed flip rate one. Under intervention, model A assigns hidden-class rates 0.5 and 3.5, whereas model B assigns rate 2 in both classes. Panel (b) shows the exact retention curves and their separation at $t = 0.6$. The structures and rates are mathematical examples, not fitted genomic mechanisms.

It vanishes for the baseline design and is strictly positive for any $t > 0$ under the specified intervention. For $t = 0.6$ it equals approximately 0.01877 nats per *independent, noiseless* endpoint observation. With n independent replicate trajectories at the same design, the joint KL divergence is n times this value. This scaling is not valid if repeated reads from one cell are incorrectly counted as independent experimental replicates. It also does not by itself imply that a given sample size guarantees successful model selection: observation noise, uncertain control effects, selection bias, and model misspecification reduce usable information.

A more general observation map, including noisy sequence, imaging, contact or inferred topology data, cannot increase distinguishability beyond the corresponding latent observations merely by applying a deterministic transformation; the data-processing inequality provides the formal limit. Additional channels can be useful precisely because they measure different aspects of the latent state rather than reformatting the same information. For a Bernoulli endpoint with smooth parameter dependence $R_\theta(t, u)$, its one-observation Fisher information is

$$\mathcal{I}_{t,u}(\theta) = \frac{\nabla_\theta R_\theta(t, u) \nabla_\theta R_\theta(t, u)^\top}{R_\theta(t, u)[1 - R_\theta(t, u)]}. \quad (98)$$

This matrix has rank at most one for a single scalar endpoint; identifying multiple parameters generally requires multiple informative observation times, controls, or measurement channels. It also requires explicit consideration of confounding and measurement calibration. The ability to design a discriminatory perturbation must not be equated with the ability to execute one safely or specifically in a biological system.

21.5 Longitudinal Genomic Measurements and Their Interpretive Boundaries

Potential evidence channels include trajectories of labeled chromatin loci, time-resolved reporter activity, single-cell multi-omics under staged perturbations, and lineage-linked measurements. Live-

cell CRISPR-based imaging can capture genomic-locus motion but remains subject to practical limitations involving labeling, spatial resolution, cellular perturbation, and accessible loci (Park and Kim, 2025). Lineage tracing can establish partial genealogical relations between sampled cells, but a lineage tree alone does not directly observe the full chromatin state history or genomic structural transformations (Short et al., 2025). Static population Hi-C measurements at several nominal times do not automatically provide within-cell contact trajectories. These different designs consequently require different likelihoods, observation-error models, and assumptions about cell identity.

For genomic two-complexes, we propose reporting an explicitly typed sequence of observations

$$D = \{(t_j, u_j, y_j^{\text{seq}}, y_j^{\text{contact}}, y_j^{\text{image}}, \mathcal{A}_j)\}_{j=0}^m, \quad (99)$$

where $\mathcal{A}_j$ records known or uncertain links between sampled cells and genomic coordinates. The notation denotes an idealized set of possible modalities; an individual experiment need not possess all channels. Inference would marginalize over compatible latent cellular states and the uncertainty of their correspondences, rather than assigning every observed change in a contact map to a unique face-attachment event. When topological summaries are used, the changing resolutions, candidate-edge support, and face rules from Sections 15–19 must be included in the model. Otherwise an apparent change in H_k may result from measurement or reconstruction choices rather than genomic dynamics.

A falsifiable computational workflow would first fit transition and emission models on held-out synthetic histories with known generators, including deliberately indistinguishable alternatives; then assess recovery under time subsampling, contact noise, lineage mismatch, and perturbation uncertainty; next compare predicted transition statistics against simple kinetic baselines and established polymer or sequence-to-function models; and only thereafter test appropriately paired biological datasets. The essential outputs are predictive distributions and calibrated uncertainties, not a single claimed reconstructed history. If no tested observable separates a detailed cellular-history model from a simpler effective model, the added topological description has not demonstrated predictive or mechanistic necessity.

21.6 Implications for the Generative-Physics Hypothesis

The present section extends our notion of an end-to-end model from a static joint law to an explicitly controlled family of path distributions. In favorable experimental regimes longitudinal and intervention-based observations may distinguish some mechanisms that share a stationary sequence distribution. They cannot, in general, establish unique microscopic causation, a group-field interaction vertex, or a spacetime condensate. Deriving $Q_\theta^{(u)}$ from an independently specified physical action, or proving a controlled kinetic approximation to that action, remains a separate unsolved task. Conversely, establishing useful effective transition constraints without such a derivation would already support the data-informed topological program. Thus the scientifically defensible hierarchy is: *measured dynamical distinctions*, *tested effective generative mechanisms*, and only then any stronger claim about microscopic field-theoretic origin.

22 History-Aware Inference on Evolving Genomic Cellular Complexes

The preceding section established that an unobserved relational state can affect the time dependence of a projected measurement even when baseline observations agree. We now allow the *cellular support itself* to change. This is a more direct finite realization of structural generation than evolving labels on the fixed cylinder of Section 3: new edges and two-cells may appear, existing cells may disappear, and the homology of successive configurations may change. The resulting

model is still a phenomenological stochastic dynamics, not a derivation of topology-changing amplitudes from the proposed group-field action. We specify the cellular operations, their algebraic admissibility, a normalized path law, observation likelihoods, and a small example admitting exact verification.

22.1 Admissible Cellular Events and Consistent Boundaries

Fix a finite ambient oriented cellular or simplicial two-complex $\mathcal{A}$, and let $\mathbf{X}_{\mathcal{A}}$ be a prescribed finite family of its subcomplexes with biological attributes. For an oriented cell σ , its cellular boundary is fixed by the orientation conventions of $\mathcal{A}$. A subcomplex contains the entire boundary of every cell it contains. In a simplicial implementation this is the usual closure under faces. Attributes, including a designated genomic contour, need not be identical across the states; any replacement of contour segments must be accompanied by a specified correspondence rather than an implicit identification of positions.

Definition 22.1 (Elementary admissible cellular event). *An event $m : \mathcal{K} \rightsquigarrow \mathcal{K}'$ is admissible when (i) both $\mathcal{K}$ and $\mathcal{K}'$ belong to the declared model family, (ii) their boundary operators obey $\partial_1^{\mathcal{K}} \partial_2^{\mathcal{K}} = 0$ and $\partial_1^{\mathcal{K}'} \partial_2^{\mathcal{K}'} = 0$, (iii) its changed cells and their incident boundaries are explicitly recorded, and (iv) any labels carried between nonidentical cells are related by a stated cellular correspondence. Under an elementary insertion, a k -cell can be added only after its boundary cells exist; under deletion, a cell with remaining higher-dimensional cofaces cannot be deleted alone. A compound edit is an ordered sequence of elementary events, or a separately defined valid replacement with a common evidence span.*

The incidence identity is a necessary structural consistency condition, but is not by itself enough to ensure biological admissibility. In particular, inserting a two-cell to fill a contact triangle is a *modeling decision* requiring a face-construction rule; three observed pairwise contacts do not prove simultaneous multi-locus interaction. The chosen rule must remain explicit in any application to Hi-C, imaging, or pangenome observations.

If $\mathcal{K} \subseteq \mathcal{K}'$ is an insertion, the inclusions of their chain groups satisfy

$$\partial_k^{\mathcal{K}'} i_k = i_{k-1} \partial_k^{\mathcal{K}}, \quad k = 1, 2. \quad (100)$$

The reverse physical operation is not a chain map pointing from $C_{\bullet}(\mathcal{K}')$ into $C_{\bullet}(\mathcal{K})$ merely because it is the reversal of the edit: the inclusion still points from the smaller complex to the larger one. For general consecutive configurations, a common subcomplex or an evidence correspondence gives a span

$$\mathcal{K}_t \longleftarrow \mathcal{Z}_{t,t+1} \longrightarrow \mathcal{K}_{t+1}. \quad (101)$$

This supports zigzag persistence or correspondence-based transport instead of treating cell deletion as an ordinary forward filtration (Carlsson and de Silva, 2010). A common intersection is available when both complexes share an ambient cellulation; inversions and segment duplications may require the more general partial correspondences of Section 17.

A dimension convention is essential. An evolving *one-dimensional* contour or contact graph can have a two-dimensional event-history support. If the instantaneous object is already a two-complex, appending a time interval or using its ordinary mapping cylinder can produce a *three-dimensional* object. The family $\{\mathcal{K}_t\}$ and its correspondence diagram therefore need not themselves form a single two-complex. The term “genomic spacetime” cannot remove this dimensional distinction.

22.2 A Normalized Law on Cellular Histories

For a finite admissible state family $\mathbf{X} = \{x\}$, let $x = (\mathcal{K}_x, \phi_x, w_x, \dots)$ contain its cellular support and relevant biochemical or symbolic attributes. An elementary event m has a target $T_m x$ and a

nonnegative hazard $r_m(x; u, \theta)$ under experimental condition u . Define the generator

$$Q_{xy}^{(u)} = \sum_{m: T_m x = y} r_m(x; u, \theta) \quad (y \neq x), \quad Q_{xx}^{(u)} = - \sum_{y \neq x} Q_{xy}^{(u)}. \quad (102)$$

For piecewise constant controls, $P_t^{(u)} = e^{tQ^{(u)}}$ is a stochastic transition matrix. The rate units are inverse physical time *only if* time and hazards have been calibrated to an experimental process. No detailed balance, fixed genome length, or equilibrium stationary law is assumed. An optional energy-based local-detailed-balance restriction would be an additional physical hypothesis, not a consequence of cellular topology.

For a completely observed event path $\omega = (x_0, (t_j, m_j)_{j=1}^J, T)$, its density with respect to the usual event-time and mark reference measure is

$$\mathbb{P}_\theta(d\omega \mid u) = \pi_0(x_0) \exp \left[- \int_0^T \lambda(X_s; u_s) ds \right] \prod_{j=1}^J r_{m_j}(X_{t_j^-}; u_{t_j^-}, \theta) dt_1 \cdots dt_J, \quad (103)$$

where $\lambda(x; u) = \sum_{m: T_m x \neq x} r_m(x; u, \theta)$. Under a specified family of event types this law is normalized by the continuous-time Markov construction. When events are hidden, the corresponding observed distribution is a sum or integral over compatible event paths; replacing it with one reconstructed history without uncertainty is generally unjustified.

Let $D_j(y_j)$ be a diagonal emission matrix with entries $E_\theta(y_j \mid x, u_j)$ at observation time t_j , and set $\Delta_j = t_j - t_{j-1}$. With a row-vector prior π_0 and a condition u_j held on $(t_{j-1}, t_j]$, the snapshot likelihood is

$$\mathcal{L}_\theta(y_{0:J} \mid u) = \pi_0 D_0(y_0) \left[\prod_{j=1}^J e^{\Delta_j Q^{(u_j)}} D_j(y_j) \right] \mathbf{1}, \quad (104)$$

where the product is ordered chronologically. This expression supports exact finite-state filtering and smoothing. For non-destructive, time-linked measurements it describes a single tracked system; independent destructive samples from different cells instead require an explicit population/lineage sampling model and must not be silently concatenated into one path.

22.3 An Exact Topology-Changing Three-State Model

Consider three loci with fixed sequence-contour edges e_{01} and e_{12} , an optional contact edge e_{20} , and an optional two-cell f_{012} with $\partial_2 f_{012} = e_{01} + e_{12} + e_{20}$. Define K_0 to contain the contour path, K_1 to include the closing contact edge, and K_2 to include the filled face as well. The inclusions $K_0 \subset K_1 \subset K_2$ are cellular. The first homology dimensions are

$$(\beta_0, \beta_1, \beta_2)(K_0) = (1, 0, 0), \quad (\beta_0, \beta_1, \beta_2)(K_1) = (1, 1, 0), \quad (\beta_0, \beta_1, \beta_2)(K_2) = (1, 0, 0). \quad (105)$$

These are topological statements about the selected cell attachments. They do not by themselves establish a biological chromatin loop in K_1 or a physical filled cavity in K_2 .

The admissible event graph and generator are

$$K_0 \xrightleftharpoons[b]{a} K_1 \xrightleftharpoons[d]{c} K_2, \quad Q = \begin{pmatrix} -a & a & 0 \\ b & -(b+c) & c \\ 0 & d & -d \end{pmatrix}, \quad a, b, c, d > 0. \quad (106)$$

The first move inserts a contact, the second attaches a face; the reverse moves delete the face and then the contact, respectively. The boundary matrices for K_0, K_1, K_2 have respective ranks $(\text{rank } \partial_1, \text{rank } \partial_2) = (2, 0), (2, 0), (2, 1)$, giving Eq. (105) by elementary cellular homology.

Proposition 22.2 (Topological observation and strong lumpability). *For the generator in Eq. (106), observe only $Y_t = \beta_1(K_{X_t}) \in \{0, 1\}$. The partition $\{\{K_0, K_2\}, \{K_1\}\}$ is strongly lumpable if and only if $a = d$. In that case the projected process is a two-state continuous-time Markov chain for every initial law, with rates $0 \rightarrow 1$ equal to a and $1 \rightarrow 0$ equal to $b + c$. When $a \neq d$, no single two-state generator reproduces the projected transition laws for every initial distribution.*

Proof. By the strong lumpability criterion (Ball and Yeo, 1993), each latent state within the observed class $Y = 0$ must have the same total transition rate to the other class. The rate is a from K_0 and d from K_2 . The $Y = 1$ class is a singleton and has total outgoing rate $b + c$. Hence the criterion is exactly $a = d$. For $a \neq d$, the small-time expansions from the two latent $Y_0 = 0$ states are

$$\mathbb{P}(Y_t = 1 \mid X_0 = K_0) = at + O(t^2), \quad \mathbb{P}(Y_t = 1 \mid X_0 = K_2) = dt + O(t^2). \quad (107)$$

They cannot be reproduced by one initial-state-independent rate from observed state 0. This leaves open the possibility of special initial laws yielding restricted Markov behavior; strong lumpability concerns all initial distributions. $\square$

Equivalently, writing $p_j(t) = \mathbb{P}(X_t = K_j)$, the expected first Betti number $\mathbb{E}\beta_1 = p_1(t)$ obeys

$$\frac{d}{dt} \mathbb{E}[\beta_1(K_{X_t})] = ap_0(t) - (b + c)p_1(t) + dp_2(t). \quad (108)$$

When $a \neq d$, this is not a closed evolution equation for $\mathbb{E}\beta_1$ alone. Consequently, a sequence of topological summary statistics is not generally a sufficient state description for forecasting structural change.

For this illustrative parameter choice, direct matrix exponentiation gives

$$\mathbb{P}(Y_{0.4} = 1 \mid X_0 = K_0) = 0.351368\dots, \quad \mathbb{P}(Y_{0.4} = 1 \mid X_0 = K_2) = 0.120195\dots \quad (109)$$

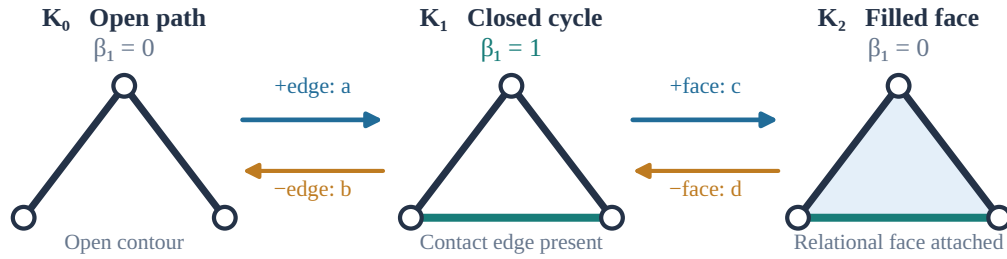
The different predictions concern conditional initial *cellular* states. Since the observed initial Betti number is identical, an experiment measuring only $\beta_1(0)$ could not choose which conditional prediction is appropriate. Additional structural observations (for example, a validated face attribute, contact data, or a controlled intervention changing event hazards) could distinguish the states, subject to measurement reliability.

22.4 Inference Across Structural Events and Multiple Genomic Resolutions

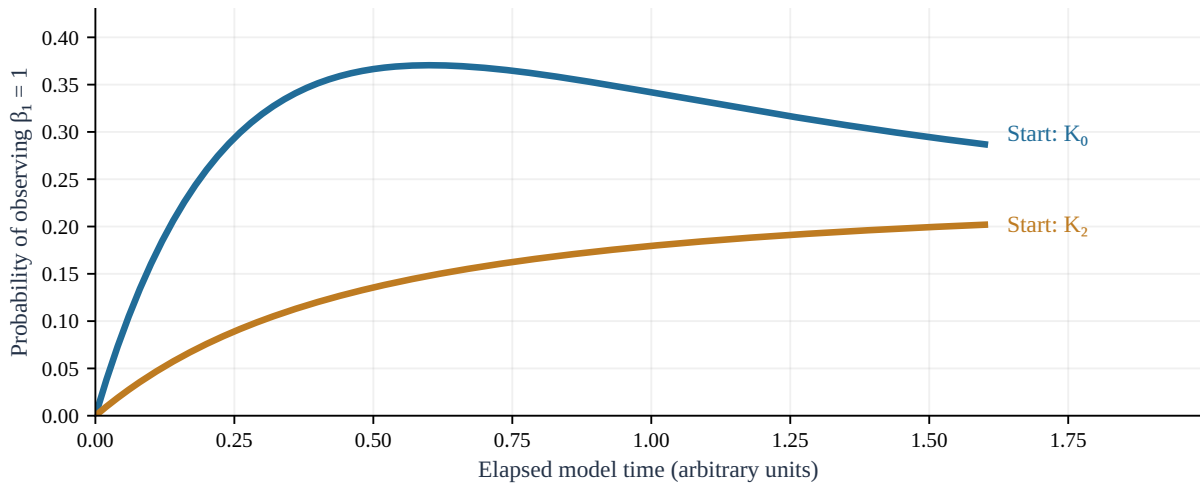
The finite example exposes the modeling distinction between a topology-changing state process and its observation. Section 23 develops a corresponding inference algorithm. For a realistic genome there are generally many admissible configurations, a rapidly growing number of possible cell attachments, and numerous candidate event histories compatible with one terminal structure. An exact transition matrix indexed by all admissible genome-scale complexes is therefore not a practical target. A feasible implementation should parameterize *local event hazards*, maintain sparse incidence lists, and approximate the posterior using constrained event simulation or particle-based path inference. Local updates can touch only cells in a modified cell's incidence neighborhood, provided the model's rate dependencies are local; long-range coupling, full topological recomputation, and particle degeneracy may substantially increase cost. Sparse matrices economize memory but do not eliminate the combinatorial number of latent complexes.

Topological observables should be updated selectively. Simple connected-component changes can sometimes be maintained by specialized dynamic-connectivity structures, whereas arbitrary deletion and reattachment of two-cells can require updates to homology reductions or persistence generators. A homology calculation at every attempted kinetic event is neither required by the

A Topology-changing cellular events



B Identical initial β_1 , distinct later observations



Illustrative rates $(a,b,c,d) = (2, 1, 1.5, 0.5)$. Matrix-exponential predictions; illustrative model.

Figure 20: A topology-changing cellular history with explicit admissible events. Panel A: adding a closing contact edge creates a first-homology class; attaching a face kills it. In both steps the existing genomic contour (black path) remains distinguished. The 2-cell is a chosen relational face, not evidence of physical cavity formation. Panel B: starting from K_0 or K_2 gives the same observed $\beta_1(0) = 0$ but distinct future probabilities of observing $\beta_1 = 1$ under the exact three-state generator with $(a, b, c, d) = (2, 1, 1.5, 0.5)$. The curves come from matrix exponentiation, not genomic measurements.

Markov definition nor necessarily feasible. Event logs can instead retain elementary edit records; selected snapshots can then be checked by exact boundary matrices and topology calculations. Across resolutions, the chain maps and information-loss diagnostics of Section 16 are required to prevent a coarse representation from appearing to preserve an event that it has actually collapsed.

If the topological state family is reconstructed from uncertain contact scores or incomplete measurements, the likelihood in Eq. (104) must be extended to a distribution over cellular constructions and correspondences. The joint posterior structure of Section 19 remains applicable, now with correlated event paths as an additional latent component. Neither comparing static Betti vectors nor naively averaging adjacency matrices retains the path uncertainty induced by multiple admissible histories.

22.5 Biological Interpretations and Testable Failure Conditions

A contact-addition event, an experimentally detected multi-locus relationship, a rearrangement of DNA sequence, and the historical face of a generative two-complex are distinct mechanisms.

They should not be assigned the same rate or the same two-cell semantics without evidence. For example, a chromatin contact may change reversibly without modifying the sequence, while a chromosome inversion changes contour correspondence and may require a new set of genomic boundary paths. Such a transformation is more naturally represented by a partial cellular span or rewriting rule than by filling a contact triangle.

Three empirical checks would be required before using the kinetic formalism as genomic inference: (i) event definitions and dwell-time models must be calibrated against measurements capable of resolving the relevant changes; (ii) topological-state observations must be compared with raw sequence/contact/imaging observations to identify information lost by projection; and (iii) held-out controlled perturbations must discriminate fitted event-rate mechanisms from simpler contact or polymer models. If correlated events, non-exponential holding times, changing cell identities, or unobserved regulatory states invalidate the Markov approximation, the state must be enlarged or the kinetic model replaced by an explicitly semi-Markov or non-Markov formulation.

The construction advances the *finite end-to-end program* by connecting lawful cellular changes to an explicit path probability and uncertain boundary/topology observations. Its fundamental limitation remains that the hazards r_m have been *specified*, not derived from a group-field action or nucleotide chemistry. A physical spacetime-condensate interpretation would require an independent microscopic origin, scale-dependent observables, and experimentally meaningful collective behavior; exact finite-state kinetics alone supplies none of those conclusions.

23 Scalable Bayesian Reconstruction of Genomic Cellular Histories

The event-based description of Section 22 supplies a normalized probability law on admissible cellular histories. Its inverse problem is different from finding a single optimal sequence of edits. Given sparse, noisy, and possibly indirect measurements, one must infer a *posterior distribution* over latent complexes and edit histories, retaining histories that agree on the observations but differ in the mechanism that generated them. The finite-state likelihood in Eq. (104) is exact, yet enumerating all admissible complexes or materializing a genome-scale generator is usually impossible. This section specifies a computational route from exact finite filtering to local event simulation and sequential Monte Carlo (SMC), identifies an important exactness guarantee and its limitations, and verifies the resulting inference logic against the three-complex construction. No chromosome-scale software implementation or empirical reconstruction is claimed.

23.1 The Posterior on Cellular Paths and Measured Observations

Let $X_t = (\mathcal{K}_t, \phi_t, w_t, \mathcal{A}_t)$ contain the changing cellular support, biochemical or symbolic attributes, and any uncertain coordinate correspondence. Let $\omega = (X_0, (\tau_\ell, m_\ell)_{\ell=1}^L)$ be the marked edit history defined by the hazards $r_m(X_t; u_t, \theta)$ of Section 22. Experimental observations are $D = (y_0, \dots, y_J)$ at ordered times $0 \leq t_0 < \dots < t_J$, with conditionally specified emission densities $g_j(y_j | X_{t_j}, \eta_j)$; η_j can contain depth, batch, modality, and alignment-error parameters. For fixed model parameters θ and calibrated emission assumptions, the history posterior is

$$p_\theta(d\omega | D, u) = \frac{p_\theta(d\omega | u)}{\mathcal{Z}_\theta(D | u)} \prod_{j=0}^J g_j(y_j | X_{t_j}(\omega), \eta_j), \quad (110)$$

$$\mathcal{Z}_\theta(D | u) = \int p_\theta(d\omega | u) \prod_{j=0}^J g_j(y_j | X_{t_j}(\omega), \eta_j).$$

The denominator is an observed-data likelihood, not the physical partition function of the static contact model. If modalities are conditionally dependent given X_{t_j} , the joint emission g_j must

retain that dependence; multiplying several marginal emissions without conditional independence would generally double-count evidence. Likewise, observations taken destructively from distinct cells do not form a trajectory of one cell merely because their collection times can be ordered. They require a hierarchical population or lineage model. Uncertainty over resolution, correspondence, and construction rules can be included by augmenting the latent state with the joint hypothesis z of Section 19; posterior averaging of incompatible incidence matrices does not create a meaningful cellular history.

The posterior may assign high mass to multiple, topologically different paths ending in the same terminal sequence. Reporting one reconstructed $\hat{\omega}$ is therefore inadequate unless the mass of competing histories and the effects of model misspecification are also reported. In particular, the observation map can be noninjective even with unlimited noiseless snapshots of some summaries, as the strong-lumpability and nonidentifiability examples of Sections 20 and 22 already establish.

23.2 Exact Filtering and Smoothing on a Finite State Family

For a finite admissible family $\mathbf{X}$ with generator Q , let $M_j = \exp((t_j - t_{j-1})Q^{(u_j)})$ and $G_j = \text{diag}(g_j(y_j | x))$. With a row-vector initial distribution π_0 , define the normalized filtering distribution $\alpha_j(x) = \mathbb{P}(X_{t_j} = x | y_{0:j})$. The recursion is

$$\alpha_0 = \frac{\pi_0 G_0}{c_0}, \quad c_0 = \pi_0 G_0 \mathbf{1}; \quad \alpha_j = \frac{\alpha_{j-1} M_j G_j}{c_j}, \quad c_j = \alpha_{j-1} M_j G_j \mathbf{1} \quad (j \geq 1). \quad (111)$$

Every denominator is positive whenever the corresponding observed data have positive likelihood. The complete observed-data likelihood factorizes exactly as

$$\mathcal{Z}_\theta(D | u) = \prod_{j=0}^J c_j. \quad (112)$$

Filtering conditions on data available up to t_j , whereas posterior smoothing conditions on *all* snapshots. Defining a backward column vector $b_J = \mathbf{1}$ and $b_j = M_{j+1} G_{j+1} b_{j+1}$ gives

$$\mathbb{P}(X_{t_j} = x | D, u) = \frac{\alpha_j(x) b_j(x)}{\sum_z \alpha_j(z) b_j(z)}. \quad (113)$$

These are standard finite-state hidden Markov recursions for a sampled continuous-time chain, now applied to structural cellular states rather than merely nucleotide identities. They are useful as an exact reference for algorithm verification, but their memory and matrix exponential costs become prohibitive when $|\mathbf{X}|$ enumerates a combinatorial family of genomic complexes. The recursions also show that a claim of *history reconstruction* requires a smoothing or path-posterior procedure; a sequence of filtering marginals does not, on its own, recover which structural edits occurred.

23.3 Event-Driven Particle Inference on Sparse Cellular States

A practical alternative represents the distribution by P weighted particles, each carrying one admissible complex together with sufficient biochemical attributes and, when smoothing is required, compressed event ancestry. The bootstrap version of SMC uses the prior kinetic law to propagate particles between observations, then reweights them by the appropriate emission and resamples. Standard particle filtering theory provides the justification for this construction (Del Moral, 2004; Doucet and Johansen, 2011); simulation of the continuous-time event path can use exact exponential waiting times for piecewise-constant hazards, in the manner of the stochastic simulation algorithm (Gillespie, 1977).

For a particle in state x under fixed experimental condition u , let $\lambda(x; u) = \sum_m r_m(x; u, \theta)$, excluding null transitions. Draw a waiting time $\tau \sim \text{Exp}(\lambda)$ and an admissible event mark with

probability r_m/λ . Apply m while preserving the required face/coface closure, orientation, and any explicit correspondence. If τ crosses the next observation or intervention boundary, propagate only to that boundary; when the control changes, update the hazards before simulating subsequent events. For the special case $\lambda = 0$, no event occurs until the model's condition changes. This is exact *conditional on the specified finite-rate kinetic model and the event sampler*, not an exact simulation of a microscopic genome physics theory.

Bootstrap filter at observation t_j . (a) Independently propagate P resampled states through the correct controlled transition kernel on $(t_{j-1}, t_j]$; (b) compute $\tilde{w}_j^{(p)} = g_j(y_j | X_{t_j}^{(p)})$; (c) record $\hat{c}_j = P^{-1} \sum_{p=1}^P \tilde{w}_j^{(p)}$; (d) normalize weights and resample particle indices with replacement; (e) retain valid event logs and parent indices for subsequent smoothing. The initial observation uses particles drawn from π_0 and the same weighting step. Equal weighting after resampling must not be confused with independent posterior draws: genealogical dependence remains. When the likelihood is sharply concentrated or rare transitions dominate, proposal kernels conditioned on upcoming observations, local rejuvenation, guided bridges, or particle Markov chain Monte Carlo may reduce degeneracy, but those variants require their own correctness arguments and complexity checks.

An inexpensive indicator of weight concentration, measured *before* resampling, is

$$\text{ESS}_j = \frac{\left(\sum_{p=1}^P \tilde{w}_j^{(p)}\right)^2}{\sum_{p=1}^P (\tilde{w}_j^{(p)})^2} \in [1, P], \quad (114)$$

when at least one weight is positive. ESS monitors instantaneous imbalance, not posterior correctness; it cannot diagnose all path degeneracy, multimodality, unidentifiable mechanisms, or a wrong face-construction rule. If all particle weights are zero, the filter has failed to represent an observation-compatible hypothesis and must change its proposal, support, or error model rather than arbitrarily normalizing a zero vector.

Proposition 23.1 (Unbiased finite-sample likelihood estimator). *Assume a fixed parameterized model with a proper initial law, exact simulation of its Markov transition kernels, finite nonnegative observation potentials with a finite normalizing likelihood, and independent unbiased resampling at every observation (e.g., multinomial resampling). For $P \geq 1$ bootstrap particles, set $\hat{\mathcal{Z}}_\theta = \prod_{j=0}^J \hat{c}_j$. Then*

$$\mathbb{E}_{\text{SMC}}[\hat{\mathcal{Z}}_\theta] = \mathcal{Z}_\theta(D | u). \quad (115)$$

This does not assert that $\log \hat{\mathcal{Z}}_\theta$ or an individual filtering or smoothing estimate is unbiased, or that finite P is adequate when observations are rare.

Proof. The unnormalized filtering measures are obtained by alternately applying the Markov kernels M_j and multiplying by the potentials g_j . For a bootstrap system with unbiased resampling, the expected empirical unnormalized measure, weighted by the product of its previous mean potentials, equals the corresponding exact unnormalized Feynman–Kac measure. This follows by induction: conditional on the pre-resampling particles, unbiased resampling preserves their normalized weighted empirical measure in expectation; propagation by the exact kernel and multiplication by the next potential then preserve the unnormalized expectation identity. Evaluating that identity against the constant function $\mathbf{1}$ at the final observation gives Eq. (115). The result is standard for bootstrap particle filters (Del Moral, 2004; Doucet and Johansen, 2011) and is stated here to specify the verification target for the cellular-history implementation. $\square$

For posterior path probabilities, retaining full ancestor chains under repeated resampling may produce severe path degeneracy: even if many particles survive at the terminal observation, they may descend from very few early histories. Fixed-lag smoothers, forward-filter/backward-simulation schemes, particle Gibbs, and local conditional proposals are candidates, provided they

represent the same defined path law. When cell identities change under insertion, inversion, or duplication, the genealogy must track the actual cell correspondences; a particle lineage index alone does not provide a mathematically meaningful correspondence between nucleotide positions.

23.4 An Exactly Checkable Filtering and Smoothing Example

We return to the three cellular states K_0, K_1, K_2 and generator in Eq. (106), using $(a, b, c, d) = (2, 1, 1.5, 0.5)$ and initial law $(1/2, 0, 1/2)$. For the illustrative binary measurement $Y_j \in \{0, 1\}$, define

$$\mathbb{P}(Y_j = 1 \mid X_{t_j} = K_0, K_1, K_2) = (0.12, 0.88, 0.16). \quad (116)$$

These likelihoods are hypothetical observation-noise choices, *not* fitted chromatin-imaging or contact-assay error rates. We observe $(y_0, \dots, y_4) = (0, 1, 0, 1, 0)$ at $(t_0, \dots, t_4) = (0, 0.35, 0.70, 1.20, 1.75)$ in model-time units.

The exact recursion in Eq. (111) yields $\mathcal{Z} = 0.0297648557\dots$, and the final filtered distribution is

$$\alpha_4 = (0.213696, 0.059496, 0.726808) \quad (\text{rounded}). \quad (117)$$

A brute-force sum over all 3^5 snapshot-state sequences reproduces the same evidence and the exact backward-smoothing marginals to floating-point precision. With $P = 1,200$ particles and 160 independent bootstrap-filter runs using seed 20261027, the mean likelihood estimate was 0.02968154 (Monte Carlo standard error 0.00009525), and the largest absolute discrepancy between the *run-averaged* filtering probabilities and the exact recursion was 0.00184. The smallest mean pre-resampling ESS fraction over observation times was about 0.506. These finite computations test the recursion and implemented sampling/weighting logic, not physical realism or scalable convergence. For numerical convenience the three-state propagation here uses the exact 3×3 matrix exponential; full-scale event-driven propagation remains proposed rather than implemented.

The distinction between filtering and smoothing is quantitatively relevant even in this tiny system. For instance, the exact probability of K_1 at $t_1 = 0.35$ is 0.64096 using only y_0, y_1 , but becomes 0.56678 when conditioned on the complete observation sequence. An interpretation of the early complex based solely on the last available snapshot would therefore lose information carried by later measurements. The full path posterior is richer still, because sequences of face attachment and removal are not determined by snapshot marginal states.

23.5 Genome-Scale Cost, Approximation Error, and Diagnostic Requirements

Let M denote the number of currently represented candidate event channels, P the particle count, J the number of observation intervals, E_j the expected number of simulated elementary edits per particle in interval j , and d_{loc} a bound on the number of hazards affected by one local edit. With an incrementally maintained indexed hazard tree, locating an event can take $O(\log M)$ and updating its local rate neighborhood can take $O(d_{\text{loc}} \log M)$; these costs are conditional on sparse local dependencies and a feasible indexed representation, not universal bounds on genomic histories. With per-observation emission cost $c_{\text{obs},j}$, a target complexity is

$$O\left(P \sum_{j=1}^J [E_j(1 + d_{\text{loc}} \log M) + c_{\text{obs},j}] + PJ\right), \quad (118)$$

plus initialization and any nonlocal topological diagnostics. Storing particles costs $O(PS_{\text{cell}})$ for per-particle sparse state storage S_{cell} , before lineage/event-log storage, which may scale with the number of recorded edits and observations. This expression concerns *computation per attempted inference run*; it says nothing about the possibly exponential number of histories, posterior concentration, or how large P must be to obtain a reliable estimate. If many hazards change after

A Cellular-history particle-filter workflow



Each move preserves cellular admissibility; observations reweight proposed histories.

B Exact filtering versus bootstrap particle estimates

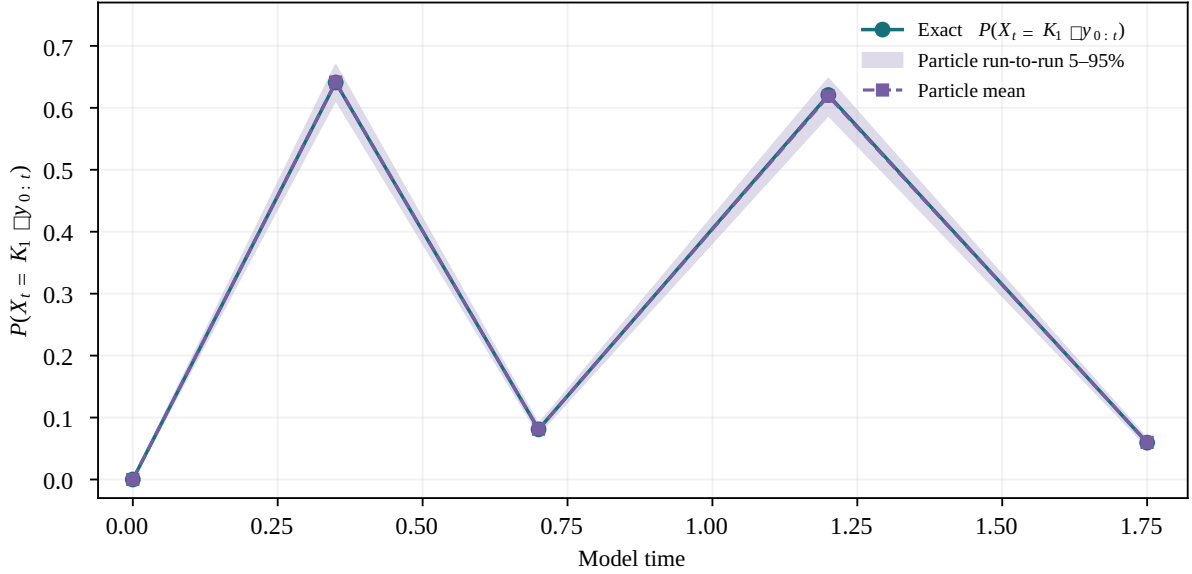


Figure 21: A finite inference example for the evolving cellular history. Panel A: the proposed local-event particle-filter workflow, with valid cellular edits, observation weighting, resampling, and retained ancestry. Panel B: exact filtered probability of the transient-cycle state K_1 at five observations (solid line), the mean of 160 bootstrap-filter runs (dashed line), and their empirical 5–95% range (shaded) for 1,200 particles per run. The toy benchmark uses exact finite matrix transition probabilities for propagation; the left-hand event-driven algorithm is a prospective scalable implementation, not a claim that a genome-scale kinetic sampler has been released.

each rearrangement, or if computing one likelihood requires dense pairwise contacts or complete persistent homology, the local-cost hypothesis fails.

A defensible implementation would therefore (i) preserve genomic-contour and face/coface constraints in every accepted event; (ii) choose a fixed physical or experimentally justified resolution for the principal analysis and propagate uncertainty over alternatives; (iii) use sparse local updates with occasional exact $\partial_1\partial_2 = 0$ and filtration checks; (iv) benchmark particle likelihoods and marginals against exhaustive finite models; (v) assess ESS, ancestry diversity, repeated-seed variability, and posterior predictive calibration; and (vi) test contact-only, topology-free, and low-dimensional kinetic baselines on withheld synthetic and biological interventions. Simulation diagnostics must include independently specified ground-truth histories at multiple sizes and noise levels. A model that fits terminal sequences but fails to recover calibrated structural events, or whose gain disappears when compared with a simple CTMC/contact model, has not established an advantage from the larger relational formalism.

The immediate engineering objective is thus a *sparse, well-validated posterior approximation*, not an exact summation over all genome-scale two-complexes. Formal particle-filter correctness

is conditional on a correct latent-state and observation model; it does not remedy a biologically unjustified face rule, mismatched cell type, or the absence of independent evidence for the proposed microscopic field action. The research program can nonetheless make empirical progress through probabilistically inferred structural histories while the stronger spacetime-condensate mechanism remains unresolved.

24 Genome-Scale Implementation Architecture and Benchmark Design

The preceding sections supply finite constructions, uncertainty bounds, and posterior recursions. They do not, by themselves, specify a feasible pipeline for an assembly-scale dataset. This section translates those mathematical objects into a proposed computational specification while separating *representation size*, *topological computation*, *stochastic-history inference*, and *biological validation*. Its estimates are derived from explicit storage assumptions, not measured runtimes of an existing genome-scale software release. Our goal is to identify algorithms and acceptance tests that would make an eventual implementation falsifiable.

24.1 Typed Genomic Evidence and Construction of an Admissible Complex

A reproducible analysis begins with a declared coordinate system: assembly version, haplotype or sample identifier, chromosome or path, interval boundaries, strand orientation, genomic resolution r , and an explicit table of missing or ambiguously mapped intervals. Let V_r denote the resulting bins. Each proposed edge carries an *edge type*, either contour adjacency, experimentally supported chromatin contact, evidence-backed structural rearrangement, or a separately declared model-derived relation. A contact weight must record its unit and processing provenance: raw count, normalized contact estimate, calibrated probability, or a transformed similarity score. These quantities cannot silently be substituted for one another or treated as transition rates.

We propose a typed evidence object

$$\mathcal{D}_r = (V_r, E_r^{\text{contour}}, E_r^{\text{contact}}, E_r^{\text{variant}}, F_r, \omega_r, \text{prov}_r, \text{mask}_r), \quad (119)$$

where F_r is a face list, ω_r holds measured or model-estimated attributes, prov_r records the construction rule and measurement origin of each cell, and mask_r marks uncertain or unavailable evidence. In particular, F_r cannot be obtained merely by labeling every visually suggestive polygon a physical three-locus interaction. The two principal alternative constructions are (i) a declared flag-complex two-skeleton generated from retained pairwise edges and (ii) a constrained two-complex with faces admitted only under a specified multiway or generative-transition criterion. They must be separately versioned because they express different scientific hypotheses.

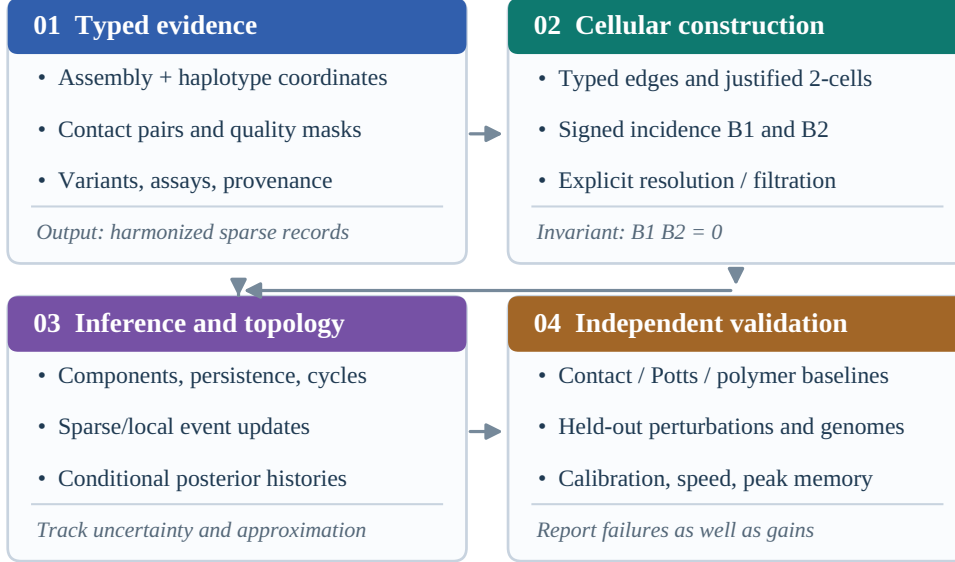
For each constructed oriented face, its boundary is an integer signed edge chain. Cell identities, signed incidences, and the attachment map are stored in a canonical order. This makes the test $B_1 B_2 = 0$ an implementation invariant, not an optional plotting diagnostic. For a linear chromosome, the contour graph is a path, not a periodically closed circle; a periodic contour is appropriate only when circularity is part of the stated biological or synthetic system. Pangenome paths require a reference-independent identifier and a separate alignment/correspondence layer, as in Section 17. Neither a sequence alignment nor a chromosome contact map alone determines the admissible 2-cells.

The proposed input interface should accept sequence and variation records with stated coordinate mappings, sparse contact matrices, optional multiway-contact observations, and measurement metadata. Cooler supplies an established sparse, multiresolution contact-matrix format, while pairtools supplies a modular read-pair processing workflow; both are preferable to inventing another incompatible primary data standard (Abdennur and Mirny, 2020; Galitsyna et al., 2024). A

conversion layer should preserve raw and transformed evidence separately. Topological calculations then consume a normalized, typed intermediate representation with an auditable sequence of filters, not a bare opaque adjacency matrix.

Genome-scale cellular inference: modular implementation

Specified pipeline • proposed large-genome workflow



Requirements: cell provenance • strict held-out target separation • fixed target + error budget • uncertainty-aware comparisons

Figure 22: Proposed genome-scale computational architecture. A typed, provenance-preserving evidence layer supports two explicitly different face-construction rules. Sparse incidence matrices and optional temporal cellular edits feed exact local checks, persistent-topology routines, and conditional history inference. The final validation layer includes matched biological baselines, calibration and compute accounting. Arrows specify proposed data dependencies, not an existing released genome-scale implementation.

24.2 Sparse Memory Budgets and the Costs Hidden by a Compact Representation

Let $n = |V_r|$, $m = |E_r|$ (including contour edges), and $f = |F_r|$. For nondegenerate oriented edges with two distinct endpoints and triangular faces with three distinct boundary edges, the signed incidence matrices $B_1 \in \mathbb{Z}^{n \times m}$ and $B_2 \in \mathbb{Z}^{m \times f}$ contain exactly $2m$ and $3f$ nonzero entries, respectively. This is a combinatorial accounting identity and says nothing about the complexity of computing persistent homology or a posterior over histories.

Proposition 24.1 (Sparse incidence storage under a declared format). *Assume the incidence coefficients are stored as signed 8-bit values, and the column-sparse (CSC) row indices and column pointers are unsigned 32-bit integers. Assume additionally that the array dimensions and number of stored entries satisfy the format’s 32-bit indexing limits. If all edges and faces satisfy the nondegeneracy conditions above, the combined storage of B_1 and B_2 , excluding allocator and metadata overhead, is*

$$\text{Bytes}(B_1, B_2) = (14m + 4) + (19f + 4) = 14m + 19f + 8. \quad (120)$$

A streaming construction from an already enumerated cell list requires $O(m + f)$ incidence writes, assuming the edge identifiers of each face are available in constant time.

Proof. A CSC matrix with z stored entries and c columns needs z one-byte values, $4z$ bytes of row indices, and $4(c + 1)$ bytes of pointers. For B_1 , $(z, c) = (2m, m)$; for B_2 , $(z, c) = (3f, f)$. Adding

the two totals gives Eq. (120). Each provided edge contributes two signed incidences and each provided triangle contributes three. If faces must first be discovered from contacts, that discovery cost is *not* included. $\square$

Table 1: Illustrative *incidence-only* memory at four resolutions. Assume a 3.1-Gbp total reference length, five unique undirected non-contour contact edges per bin as a total edge-to-vertex ratio (average non-contour degree ten; hence $m \simeq 6n$), and exactly two declared triangular faces per bin ($f = 2n$). These are synthetic budgets, not measured human-genome face counts. MB denotes 10^6 bytes. The values omit contact endpoints/weights, filtration order, data indices, H_k reduction workspace, posterior particles, provenance and file-system overhead.

Bin width	Vertices n	Edges m	$B_1 + B_2$ (MB)
100 kb	31,000	185,999	3.78
10 kb	310,000	1,859,999	37.82
1 kb	3,100,000	18,599,999	378.20
100 bp	31,000,000	185,999,999	3,782.00

These small incidence budgets must not be read as estimates of total memory. An array of contact endpoints and confidence scores, chromosome/variant annotations, a persistent-homology working reduction, and the posterior state of even a modest particle ensemble can dominate memory. Materializing one complete complex for each of P particles incurs $O(P(n + m + f))$ storage, whereas an immutable shared base plus particle-specific sparse edits can instead require $O(n + m + f + \sum_{p=1}^P |\Delta_p|)$ for the structural state, excluding ancestors and buffered observations. The latter is economical only when the number of edits per active particle stays small and the data structure supports persistent or copy-on-write updates. Keeping every full ancestral structure for J observation times can restore an $O(PJ(n + m + f))$ bottleneck. Event logs and checkpointed reconstruction provide a memory–recomputation tradeoff rather than cost-free history retention.

The number of triangular faces itself is not controlled by graph sparsity alone. For a simple graph with vertex degrees d_v , the number of triangles obeys

$$f_{\Delta} \leq \frac{1}{3} \sum_{v \in V} \binom{d_v}{2}, \quad (121)$$

with equality only under additional neighborhood-completeness conditions. More general higher-order and nontriangular cells have separately defined incidence costs. Enumerating all qualifying triangles involves neighborhood intersections and can be much more costly than writing already supplied faces. At the other extreme, if a two-skeleton is cut from a higher-dimensional flag complex, seemingly nontrivial H_2 may be an artifact of omitting three-cells, as discussed in Section 14. A storage-efficient complex can therefore still yield a misleading or computationally expensive topology.

24.3 A Resolution-Aware Algorithmic Pipeline

We propose six linked operations, each with explicit outputs and failure conditions. First, perform coordinate harmonization, replicate- and batch-aware normalization, and data-quality masking before complex construction. Second, construct typed sparse edges and faces under a declared resolution and threshold/filtration rule, recording cell provenance. Third, verify orientation and $B_1 B_2 = 0$ over integer coefficients, together with the analogous filtration-face condition that a cell appears no earlier than any boundary cell. Fourth, calculate cheap local or component-level descriptors before attempting full persistent reduction: component counts through disjoint-set union and limited-scale spectral quantities through sparse matrix-vector products are natural first checks. Fifth, where scientific questions warrant it, compute persistent cycles, selected

representatives and uncertainty-aware cross-sample comparisons with methods already validated for sparse complexes, rather than presuming a new topology engine is needed. Sixth, if the observations and model justify dynamics, infer event histories with the conditional sampler in Section 23; otherwise report static topology without inventing an unobserved transition mechanism.

In a filtered cell complex, a valid filtration value τ satisfies $\tau(\partial\sigma) \leq \tau(\sigma)$ for every incident codimension-one boundary cell. Assigning a triangle a contact-strength-based birth value earlier than any of its edges invalidates the filtration. Similarly, missing contacts should not automatically be interpreted as absent physical interactions. Different procedures for reconstructing missing contacts lead to different topological uncertainty; the perturbation bound from Section 15 only applies on the specified common cell support with fixed construction semantics.

The algorithmic complexity has three distinct regimes. Producing and storing an already specified sparse complex can be linear in its represented cell counts, subject to input parsing and index construction. Computing full persistent homology involves matrix reduction and may produce superlinear time or fill-in, even for a sparse initial boundary; the Dory human Hi-C benchmark demonstrates that particular large sparse filtrations are tractable, not that all genomic 2-complex definitions inherit that performance (Aggarwal and Periwai, 2024). Inferring cellular histories scales additionally with the number of candidate transition channels, event frequency, posterior particles and observation complexity; Eq. (118) is only a conditional *per-run* estimate. No polynomial guarantee for accurate general history reconstruction follows from the existence of a sparse complex.

A practical partitioning strategy is to use chromosome- or region-local complexes for parallel topological work, while preserving a separate layer for bona fide trans-chromosomal contacts and global rearrangements. Such partitioning is an approximation when cross-partition cells exist. Its error should be assessed by controlled inclusion of held-out cross-partition edges or by a hierarchy of overlapping windows. For structural variants, local path identifiers and signed coordinate maps should be retained across resolution changes; a bin index is not a stable biological identifier across assemblies. This principle also applies to lineage-dependent insertion and duplication histories.

24.4 Benchmark Ladder, Baselines, and Acceptance Criteria

The evaluation should proceed from mathematical correctness to biological discrimination without skipping levels. At *Level 0*, test exact incidence identities and finite rank/homology cases, gauge-quotient invariance, and the sequence-marginal equivalence of Eq. (34). These are regression tests with known ground truth; no novel biological effect may be claimed from passing them. At *Level 1*, assess static contact-derived complexes using withheld contact evidence and technical/biological replicates. Compare against a contact-graph-only baseline, distance-stratified contact summaries, and an established persistent-topology routine such as Dory. The comparison must use compatible cell and filtration semantics. At *Level 2*, examine structural perturbations and chromatin-state changes with independent measurement modalities. Compare against an appropriately calibrated polymer or MiChroM-type baseline when the target is physical contact behavior. At *Level 3*, use aligned and pangenome-reconciled assemblies to evaluate correspondence-aware structural differences with uncertainty, against accepted pangenome and genealogical representations when the target concerns variants or evolutionary history. At *Level 4*, if time-resolved lineage/perturbation data genuinely support it, evaluate inference over event histories against a low-dimensional CTMC and a topology-free particle filter.

Evaluation splits must prevent information leakage through overlapping genomic windows, shared homologous segments, related haplotypes and near-identical training assemblies. Contact data provided as input must not be scored as though predicted independently. Preprocessing, threshold selection and tuning are fitted only on the training portion, and biological controls must match the modality, cell type, resolution and sampling depth where possible. Report multiple random seeds or independent replicates, uncertainty intervals for primary metrics, peak resident

Table 2: Pre-registered reporting obligations for the proposed implementation. The table specifies criteria to report, not unperformed benchmark outcomes or universal performance thresholds.

Stage	Required comparison	Measurements and failure checks
Cell construction	Declared flag-complex rule versus evidence-constrained faces	Cell counts; unsupported-face fraction; B_1B_2 checks; stability under replicate resampling.
Static topology	Contact graph and existing persistence routines	Persistence distance with common support; localized generators; batch/coverage sensitivity; peak memory and wall time.
Structural prediction	Distance-only/contact baseline and polymer model as appropriate	Held-out predictive score; uncertainty calibration; perturbation response; input leakage audit.
Multi-genome inference	Coordinate-alignment, variant graph and tree-sequence methods matched to target	Correspondence recovery; alignment uncertainty; false matches at duplications; computational scaling.
Dynamic reconstruction	Exact small CTMC and topology-free particle filter	Event posterior calibration; held-out observations; ESS and ancestry diversity; likelihood error; cost at matched accuracy.

memory including construction, and full wall time including I/O, not merely kernel runtime. A practically useful claim needs a named target, a predefined acceptable error/calibration range and a matched cost comparison; otherwise even an attractive speedup factor or topological picture is not interpretable.

A critical negative control is the analytically integrated Potts model. For sequence-only predictions under the current finite model, it must reproduce the same observables and probability law as the augmented gauge/contact representation. Any apparent advantage of the latter on this exact subfamily indicates mismatched optimization, approximations, or errors rather than new information. Likewise, adding a face to a contact triangle changes the defined topology by construction; biological value requires independent evidence or superior out-of-sample prediction, not merely a change in β_1 . An increase in latent complexity is justified only by additional identifiable observables, better calibrated predictions, robust perturbation generalization, or a rigorously documented resource tradeoff.

24.5 Reproducibility, Software Availability, and Criteria for Progress

The proposed implementation should report a machine-readable schema version, assembly and annotation versions, all filters and cell-attachment rules, the random seeds used for posterior inference, software and hardware specifications, reference datasets or accession identifiers, and the exact definition of any learned physical parameters. Reproducible source for these procedures remains a *planned* part of the research program. The analytical finite-model checks described in this manuscript and its private exploratory verification scripts are not represented as a released, maintained genome-scale library. For project information or prospective implementations, readers may consult <https://github.com/huangwanhong-maker/The-Science-Infrastructures> or contact Wanhong Huang at huangwanhong@serendip.ngo; neither destination is asserted here to host the present paper’s complete analysis software.

The minimum criterion for leaving the purely formal stage is a reproducible real-data construction that shows (i) stable, interpretable cellular features under appropriate replicates, (ii) incremental value beyond contact graphs and existing persistent-homology methods, and (iii) documented scaling at a stated biological resolution. A stronger mechanistic contribution additionally requires prospective or held-out intervention predictions that a suitably calibrated simpler model fails to match. Even success on all these empirical targets would not prove that real genomes are

group-field condensates or that a physical local gauge symmetry has spontaneously broken. Those are separate field-theoretic claims requiring their own derivation and observables.

25 Comparison of Genomic Complexes Across Genomes and Conditions

Topological distance becomes interpretable only after declaring compatible cell-construction rules and a score-error model; Section 15 supplies a fixed-support perturbation guarantee and states the circumstances in which it does not apply.

25.1 Comparable Constructions and Coordinate Correspondence

Let D_A, D_B be datasets for two genomes, two homologous regions, or one genome measured under distinct cell conditions. A comparison begins with parallel specifications of resolution, normalization, confidence thresholds, face rules, cell types, and missing-data handling. In particular, homologous positions, syntenic blocks or paths through a pangenome graph must be related by a defensible correspondence α , possibly partial or uncertain. Separate experiments using different chromatin assays, cell types, or sequencing depth cannot be interpreted as genomic differences without appropriate adjustment. Comparisons across species additionally require accounting for rearrangements, repeats, duplications and uncertain orthology.

Given compatible filtrations, a first summary is $\beta(\mathcal{K}_A)$ versus $\beta(\mathcal{K}_B)$, followed by persistence diagrams D_k^A, D_k^B and a selected dissimilarity such as

$$d_{\text{PH}}(A, B) = \sum_{k=0}^2 a_k d_B(D_k^A, D_k^B), \quad a_k \geq 0. \quad (122)$$

Here d_B denotes bottleneck distance with a shared parameter scale and a declared convention for essential features. The expression is a user-chosen descriptor, not automatically an evolutionary metric or a complete invariant. Stability bounds for persistence diagrams apply under particular assumptions on the underlying filtration functions and domain; they do *not* provide a general bound for separately reconstructed, unaligned genomes (Cohen-Steiner et al., 2007). Interval-resolved cycle localization, effect sizes, replicate variation, and biological covariates are required to interpret observed differences.

In Figure 23, the left complex has four vertices, four edges and no two-cells, with rank $B_1 = 3$, hence $\beta_1 = 4 - 3 = 1$. In the right complex, the added diagonal produces two independent triangular boundaries under the stated flag rule, so there are five edges, rank $B_1 = 3$, and rank $B_2 = 2$. Thus $\beta_1 = 5 - 3 - 2 = 0$. This is an exact algebraic example of a *construction-dependent* topological difference, not a measured biological result.

25.2 Transformations, Homology Maps, and Time-Varying Complexes

If a verified cellular map $f : \mathcal{K}_A \rightarrow \mathcal{K}_B$ is available, it induces chain maps $f_{\#,k}$ obeying $B_k^B f_{\#,k} = f_{\#,k-1} B_k^A$ and hence maps on homology $f_* : H_k(\mathcal{K}_A) \rightarrow H_k(\mathcal{K}_B)$. This identity is an important algebraic consistency requirement for any proposed transformation representation. An approximate sequence alignment or a graph correspondence, however, is not automatically a cellular map; faces and orientations must also be respected, or the map must be replaced by an explicit correspondence or common refinement.

For an insertion, deletion, inversion, recombination event, or contact-remodeling event, a topology-changing history might be encoded by a span $\mathcal{K}_t \leftarrow \mathcal{L}_t \rightarrow \mathcal{K}_{t+1}$ representing preserved shared cells. When a sequence of complexes has both additions and removals, zigzag persistence provides an appropriate formal framework; it does not demand monotonically growing filtrations

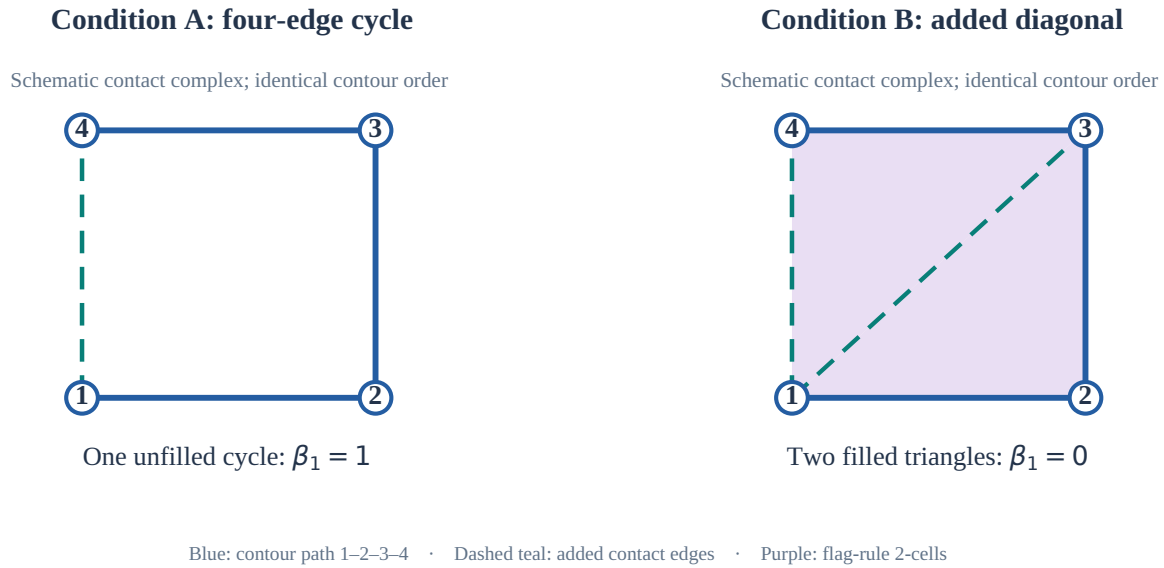


Figure 23: Comparison of two hypothetical genomic conditions at matched contour coordinates. The observed contact graphs (left) can yield distinct filled two-cells under one specified construction rule (middle), and hence different persistence signatures or local cycle representatives (right). A displayed unfilled cycle is a feature of the mathematical construction; biological interpretation requires controls, assay provenance, and perturbation evidence. This is an explanatory example, not a result from genomic datasets.

(Carlsson and de Silva, 2010). A one-parameter filtration describes scale, whereas this span describes comparisons or changes across snapshots. Neither parameter should automatically be identified with physical elapsed time or genomic contour distance. The empirical order of events and causal mechanisms require additional data.

An analysis can ask which topological signatures remain robust under (i) changes in threshold and resolution, (ii) verified cell maps across homologous regions, (iii) transformations inferred from ancestral histories, and (iv) independently measured physical perturbations. These are distinct notions of preservation. Genomic variants could change sequence attributes without changing the chosen topology, while altered chromatin contacts can change topology without a DNA sequence change. A joint analysis must accommodate both cases.

25.3 Biological Tasks and Nontrivial Baselines

Possible applications include detecting structural changes linked to regulatory remodeling, comparing cell states, relating structural variants to changes in interaction topology, and testing whether conserved syntenic neighborhoods retain selected algebraic descriptors. No specific disease, regulatory effect or evolutionary mechanism follows from Betti-number changes alone. Existing methods include pangenome alignment, tree sequences and ancestral recombination graphs for sequence history; polymer simulations for chromatin; and genome-scale TDA for contact structures (Hickey et al., 2023; Kelleher et al., 2019; Di Pierro et al., 2016; Aggarwal and Periwai, 2024; Chiang and Wu, 2026). These must remain reference methods. The proposed cellular approach adds scientific value only when its explicit faces, attributes, transformation maps or cross-scale conditions produce independently testable insight beyond those alternatives.

26 An Algebraic Description of Attributed Genomic Relational Structure

26.1 Cellular Chains, Cohomology Rings, and Boundary Constraints

A direct algebraic reformulation of the genomic complex consists of the chain data

$$\mathfrak{A}(D) = (C_2 \xrightarrow{B_2} C_1 \xrightarrow{B_1} C_0, \omega, a, \mathcal{F}, \mathcal{G}), \quad B_1 B_2 = 0, \quad (123)$$

where cellular attributes and coefficient systems are optional and $\mathcal{G}$ denotes a separately justified symmetry action. In the cochain dual, $C^k = \text{Hom}(C_k, \mathbb{F})$ and $\delta^k = (B_{k+1})^\top$ in compatible bases, so $\delta^{k+1}\delta^k = 0$. Under an appropriate cellular or simplicial diagonal approximation, a cup product induces the cohomology ring $H^\bullet(\mathcal{K}; \mathbb{F})$. This ring can carry multiplicative information not captured by Betti numbers, although a degree-one product into H^2 is necessarily trivial for the existing cylinder because $H^2 = 0$. The ring is still only a topological descriptor of the *constructed* complex, not a chemical invariant of a sequence independent of its construction.

For filtration-based comparisons, one may study persistence modules and morphisms; for explicit transformations, chain maps and mapping cones characterize induced differences. Torsion over $\mathbb{Z}$ and cup products are supplementary choices with different complexity and interpretive costs. The objective is a family of *auditable mathematical objects*, rather than one opaque composite number that purports to identify the biological function of a genome.

26.2 Cellular Sheaves and Local-to-Global Compatibility

To represent heterogeneous local information, assign a vector space $\mathcal{F}(\sigma)$ to each cell σ and a linear restriction/extension map $r_{\sigma\tau} : \mathcal{F}(\sigma) \rightarrow \mathcal{F}(\tau)$ for each incidence $\sigma \leq \tau$, satisfying composition consistency along chains of inclusions. Under this convention the cellular coboundary is

$$(\delta_{\mathcal{F}}^k x)_\tau = \sum_{\sigma^k \prec \tau^{k+1}} [\tau : \sigma] r_{\sigma\tau} x_\sigma. \quad (124)$$

When the maps satisfy the cellular sheaf compatibility conditions, $\delta_{\mathcal{F}}^{k+1}\delta_{\mathcal{F}}^k = 0$. Cohomology and sheaf Laplacians then quantify algebraic compatibility of locally assigned quantities. For a pair of incident vertices, compatibility means their features agree after mapping into an appropriate common edge stalk, not that their raw dimensions or values must coincide. Biologically justified maps are the main challenge: fitting arbitrary maps to labels may improve prediction without implying a discovered physical law.

Sheaf neural networks have been proposed as a generalization of graph diffusion for heterogeneous relational signals (Hansen and Gebhart, 2020). An attributed genomic sheaf could provisionally assign sequence features, regulatory measurements, or spatial-state embeddings to different cell types, then learn or constrain their incidence maps. Multi-locus readouts and confidence calibration would need to be handled explicitly. The constant sheaf recovers ordinary cellular cohomology; a nonconstant sheaf is a richer, *different* model whose predictions must be evaluated against simpler graph networks.

26.3 Symmetry and Holonomy as Optional Additional Data

A group-valued label $U_e \in G$ on an oriented edge has the transformation law $U_{ij} \mapsto g_i U_{ij} g_j^{-1}$ under a chosen gauge action. A closed directed path based at i has ordered holonomy U_γ transforming by conjugation $U_\gamma \mapsto g_i U_\gamma g_i^{-1}$, and a class function such as $\chi_\rho(U_\gamma)$ is consequently invariant. These are mathematically valid gauge-invariant descriptors of a specified connection. Their use on a genomic contact complex requires independently motivated edge variables and physically

or biologically interpretable transformation laws. The earlier $\mathbb{Z}_2$ finite gauge lift is an exact redundant parametrization; repackaging it as holonomies cannot create new observable physics. Gauge structure is thus optional in Eq. (123), while ordinary chain algebra remains useful without it.

27 Learning Genomic Complexes and Effective Dynamics with AI

27.1 Distinct Roles for Language Models and Genomic Foundation Models

General-purpose large language models can assist with literature synthesis, expressing cell-construction rules, preparing computational experiments, checking notation and generating candidate proofs or code. They do not establish incidence identities, characterize experimental mechanisms, or certify theorems simply by producing plausible explanations. Suggested identities should be checked using symbolic algebra or exact integer arithmetic; stochastic results need seeded replication, baselines and data provenance. The central value is accelerated *research workflow*, subject to external verification.

A separate role is played by trainable sequence foundation models, which can produce features $z_i = f_\eta(w_{i-r:i+r}, \mathcal{A})$ for a window around a locus. Such features can enter a relation proposal model, but their context length, pretraining corpus, task leakage and calibration must be assessed independently. Existing sequence-to-function and genome foundation models offer strong baselines (Avsec et al., 2026; Brixi et al., 2026); no new field dynamics follow merely from embedding a sequence with a language model.

27.2 Constrained Structure Learning with Uncertainty

Let D contain sequence, contact, multi-locus and functional observations and $\mathcal{R}$ the explicit admissibility rules for cells. A proposed inference system could approximate

$$Q_\theta(\mathcal{K}, \mathcal{F}, \phi \mid D, \mathcal{R}), \quad (125)$$

with hard support only on valid cellular structures and consistent sheaf maps. Training can combine likelihood or predictive losses with biologically selected consistency tests and complexity regularization. Rather than independently guessing arbitrary B_1 and B_2 , the generator should construct vertices, oriented edges and permissible face-attaching maps, automatically maintaining $B_1 B_2 = 0$. If a neural architecture outputs an arbitrary dense incidence array and merely penalizes $\|B_1 B_2\|$, it may still produce invalid discrete complexes at finite penalty. Conditioning on valid combinatorial constructions is safer.

Uncertainty over edges, faces, homology generators and scale-dependent features matters. Because persistence changes discontinuously when filtrations alter critical ordering, surrogate differentiable topological objectives do not obviate exact post-training verification. Attribution of a learned face to a physical multiway process requires independent observational or perturbational support. An honest output is a distribution over candidate complexes, not a unique “true genome spacetime” reconstructed from insufficient measurements.

27.3 Learning an Effective Field Model on a Constructed Complex

A tractable intermediate model could place a field $\phi \in C^k(\mathcal{K}; \mathbb{R}^p)$ on a data-derived complex and define

$$E_\theta(\phi \mid \mathcal{K}, D) = \frac{1}{2} \langle \phi, L_k(\mathcal{K}) \phi \rangle + V_\theta(\phi; \mathcal{K}, D), \quad (126)$$

with normalized $P_\theta(\phi \mid \mathcal{K}, D) \propto \exp[-\beta E_\theta]$ only when the associated partition function is finite. One might fit an effective stochastic generator instead of an equilibrium law for a driven biological system. These are learned *effective descriptions* of selected observables; they are not deductions of the original GFT action, and a fit does not confer physical uniqueness or causal identifiability. The added layers should be evaluated through held-out contact perturbations, sequence interventions where experimentally available, uncertainty calibration, and ablations against a pairwise Potts/contact model and ordinary graph neural networks.

28 Quantum Algorithms and the Conditional Prospects for Genomic Topology

28.1 Boundary Operators and Quantum Linear-Algebra Subroutines

The sparse boundary matrices B_k and Hodge Laplacians L_k in Eqs. (40)–(42) are natural candidates for certain quantum linear-algebra primitives. Quantum algorithms for estimating persistent Betti numbers or spectral properties have been developed, but their complexity depends on how simplices and boundary operators are accessed, the relevant spectral gap, approximation guarantee and cost of extracting a classical answer (Berry et al., 2024; McArdle et al., 2026). In particular, obtaining a *normalized* dimension or spectral density can be a different task from enumerating every representative cycle with biological locus annotations.

A 2026 complexity-theoretic result identifies a persistence-related problem with a provable quantum speedup under its formal access and promise conditions (Gyurik et al., 2026). This does not establish a speedup for real genome data in our construction: preprocessing contact observations, preparing input states, handling noise, executing error-corrected circuits and reading interpretable outputs may dominate the putative computational advantage. Classical sparse homology and streaming methods remain the practical baseline, particularly for low-dimensional, sparse complexes with biologically required localization.

28.2 Topological Quantum Hardware and Genomic Topological Computation

Topological quantum computing is a hardware/information-encoding paradigm that seeks topological protection for quantum logical operations. It must be distinguished from *topological data analysis*, which studies topological features of classical or quantum-provided datasets. Topologically protected qubits do not automatically accelerate homology calculations. Future fault-tolerant quantum hardware may make certain spectral, persistence or constrained-sampling primitives feasible if appropriate oracles and advantage regimes exist. This is a conditional research direction, not an operational component of our present genomic analysis.

We propose a staged criterion: first match optimized classical algorithms on public fixed filtrations; next specify a quantum task with complete input-oracle and output-precision assumptions; then estimate end-to-end logical and physical resources; finally test whether the desired *biological* result survives any reduced quantum readout. Absent these steps, quantum complexity claims would be premature. In particular, the original group-field-theory inspiration is not a reason to presume that quantum computation is inherently required to analyze an observed genomic 2-complex.

29 A Falsifiable Program for Genome-Scale Topological Analysis

29.1 An Implementation Sequence Independent of the Microscopic Action

The most feasible first study would select one high-quality publicly documented chromatin-contact dataset at a defined resolution, construct a sparse contour-plus-contact graph, attach faces using separately labeled flag and multiway-evidence rules, and compute homology and persistent features with established software. Each step should record the data version, coordinate system, filtration, edge/face counts, chain identities, random seeds where relevant, CPU/GPU memory and runtime. An independent contact-based baseline should be run with identical train–test splits. The paper currently supplies mathematical constructions and exploratory finite-model checks, *not* a released end-to-end implementation of this program.

The next experiments would compare matched replicates, perturbations of chromatin contacts, and selected orthologous or pangenome-aligned regions. A direct multi-condition study could test whether localized persistent cycles or sheaf-compatibility scores improve prediction of held-out experimentally observed regulation after controlling for contact strength, genomic distance, known TAD/loop callers, and local sequence embeddings. Adding faces or a richer algebraic coefficient system must be justified by a measurable gain. A null result would be scientifically informative: it would indicate that these higher-order structures add no demonstrated predictive value for that task.

Table 3: Research questions, currently established methods, and falsifiable prospective contributions of the proposed genomic cellular framework. All entries in the final column are research objectives, not demonstrated advantages.

Research area	Existing methods	Potential additional contribution
Chromatin topology	Hi-C-based persistent homology (Dory, topoHiC), graph and polymer models	Explicitly attributed cellular faces, localized classes, matched-condition perturbation analysis
Evolutionary and structural variation	Tree sequences, ancestral recombination graphs, pangenome paths and structural-variant calls	Tested maps between changing complexes, uncertainty over transformation-related invariants
Cross-scale regulation	Sequence-to-function predictors, multiscale contact maps and polymer models	Joint constraints on sequence, contacts and scale-dependent relational features
Topological algebra	Simplicial persistence, homology, graph spectra, cellular/sheaf models	Biologically justified sheaf restriction maps and interpretable cohomological signatures
Computational learning	Genomic foundation models, graph/cellular neural networks	Valid-cell structure learning with calibrated topology uncertainty and physical observables
Quantum extensions	Quantum TDA and sparse spectral algorithms (research-stage)	A conditional improvement on a narrowly specified, resource-audited invariant task

29.2 The Contribution Criterion and its Relationship to Generative Physics

The long-term field-theoretic hypothesis and the near-term topological pipeline have distinct proof obligations. In the latter, formal success means valid chain operators, defensible face semantics, reproducible topology, and superior or complementary biological explanation under well-controlled comparisons. In the former, formal success additionally demands a microscopic dynamical derivation, physically motivated symmetries, and a controlled collective limit. Observational topology could help choose effective interactions and constrain future action functionals, but cannot uniquely determine S from homology or persistence diagrams. The ability to compute a large genomic complex is a practical achievement, yet it is not by itself evidence for gauge

symmetry breaking, a spacetime condensate, or an original theory of DNA chemistry.

30 Existing Genome Methods and Testable Contributions

The genome-scale topological analysis literature itself is a required baseline: Dory and topoHiC already compute persistent features in Hi-C-derived genomic structures (Aggarwal and Periwai, 2024; Chiang and Wu, 2026). The present program accordingly tests added cellular attributes, face provenance, cross-sample maps and sheaf consistency rather than presenting homology on contact data as a new algorithm in isolation.

30.1 Established Models and Their Demonstrated Capabilities

The proposed formalism enters a field already containing mature large-scale methods. Sequence foundation models address long-context prediction and generative modeling: Evo 2 was trained on trillions of genomic nucleotides with a million-token context (Brix et al., 2026), while AlphaGenome processes one-megabase inputs to predict multiple regulatory modalities, including contact-related outputs and variant effects (Avsec et al., 2026). These capabilities should not be recast as functions newly enabled by a geometric representation.

Graph representations also have substantial prior art. Minigraph-Cactus constructs pangenome graphs from whole-genome assemblies, demonstrated at the scale of 90 human haplotypes (Hickey et al., 2023). In population genetics, tree sequences and ancestral recombination graphs represent the partial genealogical histories of correlated genomes; `tsinfer` and `tsdate` support scalable inference and chronological estimation in this setting (Wohns et al., 2022; Kelleher et al., 2019). These objects are not the same as the present cellular history two-complex, but both domains already formalize evolution through structured graph-like histories.

In chromatin physics, MiChroM introduced transferable, maximum-entropy-derived polymer interaction models for chromosome organization (Di Pierro et al., 2016); OpenMiChroM offers computational implementations using modern molecular simulation software (OpenMiChroM Developers, 2025). Such work models physical interactions and experimentally measured contact patterns, establishing an especially demanding baseline for any claimed mechanism-to-structure contribution. Their accomplishments also suggest opportunities to integrate a symbolic boundary readout with physical dynamics, conditional on avoiding duplicate modeling of existing effects.

30.2 Mechanistic Interpretation and Identifiability

A history-aware model would be scientifically informative if it distinguished physically plausible paths that produce the same observed sequence, or if its latent constraints made intervention predictions that an equally calibrated simpler model could not. It should therefore output uncertainty over histories instead of selecting a unique explanation from an underdetermined observation. Additional time-resolved experiments, contact measurements, evolutionary data, and controlled perturbations would be needed to identify model components. A fitted statistical association is not, by itself, evidence that a particular latent 2-cell or gauge variable is an observable molecular process.

The exact Potts reduction is an important negative control. On sequence-only targets, a baseline implementing Eq. (34) should agree with the full finite model up to sampling or numerical error. A gauge-augmented version cannot legitimately claim better predictive capacity from the redundant variables alone. New explanatory value requires added identifiable physical variables, transition kernels, additional observables, or a measured reduction in cost from a valid algorithmic optimization. The latter must be established on the same hardware, dataset, and accuracy target rather than inferred from representation-theoretic parallels.

Table 4: Established genomic analysis approaches and prospective, unverified contributions of a generative relational framework. The rightmost column specifies research questions rather than demonstrated performance improvements.

Research area	Established methods	Potential contribution to investigate
Evolutionary history	Ancestral recombination graphs; succinct tree sequences; <code>tsinfer/tsdate</code>	Integration of evolutionary ancestry with independently modeled physical and regulatory transformations; posterior uncertainty over jointly constrained histories.
Chromatin physics	Polymer and statistical-mechanical models; MiChroM; OpenMiChroM	Coupling physically calibrated interactions to explicit symbolic boundary observables and state-dependent generative transitions.
Structural variation	Pangenome graphs, genome aligners, structural-variant callers	Modeling sequences of rearrangement events and their uncertainty, rather than merely representing alternative sequence paths.
Multiscale regulation	AlphaGenome and other long-context sequence-to-function models	Derivation and testing of scale-dependent effective constraints, cross-resolution compatibility, and responses to specified physical perturbations.

31 Benchmarks for Scalable Generative Relational Inference

31.1 A Testable Progression of Tasks

We propose a progression with increasing scientific demands. The first task is algorithmic equivalence: independently implement the collapsed Potts sequence law and the original augmented law, verify agreement on small exactly enumerable systems, and measure computational cost on synthetic sparse graphs of increasing N and M . The second task is conditional prediction on observed chromatin contacts or regulatory measurements for appropriately split genomic regions, comparing a simple distance-only contact baseline, a sparse Potts or factor-graph baseline, and calibrated polymer-physics models. The third task is generalization to held-out structural perturbations, such as experimentally controlled changes of binding-site organization or local chromatin structure. The fourth task is multi-genome or multi-environment modeling with independent genealogical and variation-graph references.

The evaluation quantities should include held-out likelihood (where defined), calibrated predictive uncertainty, intervention-response error, structural/contact accuracy, peak memory, sustained throughput, and wall-clock time at matched error levels. At scale, model building and data loading must be accounted for along with steady-state inference. Statistical comparison must avoid leakage through homologous or overlapping windows, related individuals, or variants derived from the same training reference. When physical observables are the evaluation target, an externally supplied reference contact map should not be counted as an independently predicted contact map.

31.2 Ablations That Can Falsify the Claimed Contribution

Four distinct ablations are necessary: remove the 2-complex history while retaining the boundary factor model; remove or collapse gauge-redundant variables while preserving all observables; replace calibrated physical interactions with equally expressive statistical couplings; and remove the coarse-scale consistency constraints. If the richer model does not improve a predefined metric, identify a recoverable mechanism, or justify a computational tradeoff on an appropriate dataset, then its purported advantage in that setting is unsupported. No result from the present manuscript meets this benchmark yet.

For future work, the defensible contribution is the prospect of making *which transformations*

occur and what physical constraints permit them part of a joint inference problem. The current paper instead provides a mathematically checked finite prototype and an explicit list of conditions under which that prospect can be evaluated. The distinction between those achievements should remain visible when interpreting the title's condensate and symmetry-breaking language.

32 Connecting Generative Physics to Genomic Topological Inference

32.1 A Common Structure with Two Distinct Directions of Explanation

The mathematical developments above support two different research directions that must not be conflated. In a *physics-to-observation* direction, a specified microscopic or effective dynamics assigns weights to histories, histories induce cellular structures, and physical observables determine sequences, contacts, or regulatory measurements. In a *data-to-structure* direction, observed genomic and contact measurements are used to construct admissible cellular complexes and infer their topological descriptors. The first direction attempts to *explain the production* of a structure; the second *characterizes a representation* of measurements under declared construction rules. Neither direction logically entails the other.

The first direction would require a model-dependent succession of maps and stochastic kernels,

$$(\Omega_{\text{mic}}, P_{\vartheta}^e) \xrightarrow{Q_{\vartheta}} (\mathcal{H}, \mathcal{K}, X) \xrightarrow{R_{\vartheta}} (w, C, Y) \xrightarrow{A_{\ell}} (\mathcal{K}_{\ell}, T_{\ell}), \quad (127)$$

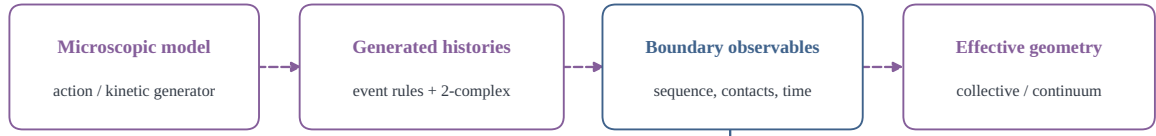
where Ω_{mic} denotes a *hypothetical* microscopic state or path space, Q_{ϑ} would induce an admissible history/complex law, R_{ϑ} is a specified observation channel, and A_{ℓ} describes the measurement-to-complex construction and its analysis at resolution ℓ . Here C denotes contact evidence, Y additional measured endpoints, and T_{ℓ} a selected topological descriptor. The arrow to A_{ℓ} is schematic: for predictive evaluation, A_{ℓ} is restricted to admissible inputs and may not use the held-out target Y . These symbols are *typed roles*, not a claim that all four arrows have been derived. In particular, A_{ℓ} may depend on detection thresholds, missing-data rules, genomic coordinates, and higher-order cell semantics. It is generally a data-analysis procedure, not a physical time evolution operator.

The second direction takes permitted input evidence $D_{\text{in}} = (w, C, \text{allowed annotations}, \dots)$ and defines an admissible family $\{\mathcal{K}_{\ell}(D_{\text{in}}, \gamma) : \gamma \in \Gamma_{\ell}\}$ and uncertainty over cell-building choices γ . A withheld endpoint Y must not enter D_{in} , feature selection, threshold tuning, or face attachment when Y is the prediction target; its use there would create target leakage. It may proceed without a microscopic path integral. The cellular incidence relation $B_1 B_2 = 0$ constrains its output, but does not identify the processes that produced D_{in} . Conversely, a hypothetical group-field amplitude for complexes does not automatically determine the biological rules used to construct faces from Hi-C observations. Figure 24 distinguishes these directions and their empirical points of contact.

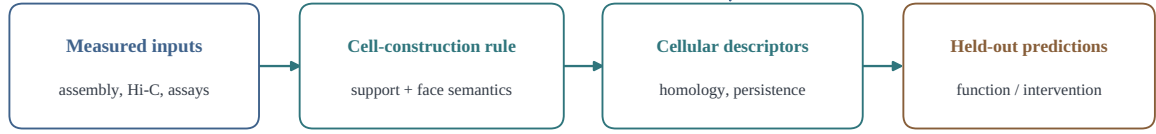
32.2 Probability Measures and the Limits of Microscopic Reconstruction

A physical measure is not supplied merely by writing a field action. A real nonnegative finite statistical weight can be normalized on a specified reference measure, but an oscillatory quantum amplitude $\exp(iS/\hbar)$ is not itself a classical probability measure. If an underlying quantum group-field model is proposed, the state, gauge treatment, observable algebra, and rule for assigning probabilities to measurable events must be supplied before Eq. (127) acquires statistical meaning. A gauge-variant condensate expectation value alone is not a physical order parameter. The separately tested four-state gauge lift in Section 5 is precisely redundant after quotienting and does not resolve these microscopic questions.

PHYSICS-TO-OBSERVATION: DERIVATION STILL REQUIRED



DATA-TO-STRUCTURE: EXPLICIT MATHEMATICAL CONSTRUCTIONS



EVIDENCE GATES

- (1) Incidence & filtration validity (2) Matched predictive baselines (3) Intervention-specific kinetics
- (4) Gauge-invariant scaling diagnostics for a genuinely derived collective phase

Figure 24: Two research directions and their evidential status. Dashed arrows indicate microscopic generativity and condensate mechanisms that remain to be derived. Solid arrows denote mathematically specified observation, cellular construction, and inference operations, subject to their stated model assumptions. Experimental interventions and independent held-out predictions connect the directions; matching a reconstructed topology alone cannot establish a microscopic origin.

For a finite or suitably measurable classical model, let $P_{\mathcal{I}}^e$ be a normalized probability measure on latent histories Ω , and let $M : \Omega \rightarrow \mathcal{D}$ be a deterministic observation map. Its observable law is the pushforward

$$P_{\mathcal{I}}^{e,\text{obs}}(A) = P_{\mathcal{I}}^e(M^{-1}(A)), \quad A \subseteq \mathcal{D}. \quad (128)$$

A stochastic observation kernel is handled by $P_{\mathcal{I}}^{e,\text{obs}}(A) = \int P_{\mathcal{I}}^e(d\omega) R(A | \omega, e)$. The latter form is appropriate when measured contacts are noisy, missing or sampled. It remains necessary to distinguish $P(\text{observation} | \text{latent state})$ from $P(\text{latent state} | \text{observation})$; changing one into the other requires the prior and a normalizing evidence term.

Proposition 32.1 (Observation maps cannot increase relative entropy). *Let P, Q be probability distributions on a finite latent history space with P absolutely continuous with respect to Q . Let R be the same stochastic observation channel under both candidate models. If PR and QR denote the resulting observable distributions, then*

$$D_{\text{KL}}(PR || QR) \leq D_{\text{KL}}(P || Q). \quad (129)$$

In particular, if $PR = QR$, the models are observationally indistinguishable through R , regardless of differences in the latent history distributions.

Proof. The finite log-sum inequality, applied for each observed value y to the nonnegative terms $P(\omega)R(y | \omega)$ and $Q(\omega)R(y | \omega)$, gives $(PR)(y) \log[(PR)(y)/(QR)(y)] \leq \sum_{\omega} P(\omega)R(y | \omega) \log[P(\omega)/Q(\omega)]$. Summing over y and using $\sum_y R(y | \omega) = 1$ proves the inequality. The last assertion follows from the equality of the observable laws. The analogous measurable-space result requires the standard absolute-continuity hypotheses. $\square$

A simple example shows that even an exact topological readout may discard all evidence about a generative difference. Let $\Omega = \{h_1, h_2, h_3, h_4\}$, with

$$P = (0.4, 0.1, 0.4, 0.1), \quad Q = (0.1, 0.4, 0.1, 0.4). \quad (130)$$

Suppose a coarse topological observation labels $\{h_1, h_2\}$ with 0 and $\{h_3, h_4\}$ with 1. Both models assign $(0.5, 0.5)$ to this observation, despite $D_{\text{KL}}(P\|Q) = 0.6 \log 4 \approx 0.83178$ nats. An additional validated measurement distinguishing $\{h_1, h_3\}$ from $\{h_2, h_4\}$ instead produces marginal laws $(0.8, 0.2)$ and $(0.2, 0.8)$, which are distinguishable. The example is algebraic and is not a model of an observed genome. It demonstrates that collecting *different* evidence can be more valuable than replacing a lost observation with an elaborate latent reconstruction.

The analogous difficulty applies to inverse group-field inference. Different microscopic actions or interaction kernels could induce the same law over measured boundary strings and contacts; recovering a fitted effective coupling would not identify a unique fundamental action. Section 20 proves a concrete instance of this issue within the contact model, and Sections 21–22 show how explicitly specified interventions and event histories can sometimes separate otherwise equivalent mechanisms.

32.3 A Kinetic Bridge for Changing Cellular Structures

If the goal is to analyze biological processes rather than formal Monte Carlo histories, admissible edge, face and sequence operations must carry empirically meaningful rates. Consider a finite event-driven system whose state X_t contains a valid cell complex, its attributes, and the environmental variables retained in the model. For each applicable event r , let $\lambda_r(X_t; e_t) \geq 0$ be a rate and define the total escape rate $\Lambda(X_t; e_t) = \sum_r \lambda_r(X_t; e_t)$. With known initial distribution p_0 and a nonexplosive time-inhomogeneous jump process, a specified history of jump times $0 < t_1 < \dots < t_n < T$ and event types $r_1, \dots, r_n$ has path-density factor

$$p_0(X_0) \exp \left[- \int_0^T \Lambda(X_t; e_t) dt \right] \prod_{j=1}^n \lambda_{r_j}(X_{t_j^-}; e_{t_j}). \quad (131)$$

The complete history probability is obtained using the associated event-time reference measure and summing over event sequences. Each permitted event must update incidence data consistently, retaining $\partial_1 \partial_2 = 0$; an insertion or duplication may change boundary cardinality, while a contact update need not change homology. Equation (131) specifies a phenomenological stochastic dynamics once rates are provided. It is *not* a derivation of those rates from the group-field action of Section 8.

A proposed physical derivation would have to show how a calibrated microscopic generator or an appropriate limit determines both the accessible event types and their rates. In particular, active chromatin remodeling, replication and repair need not obey equilibrium detailed balance. Integrating out microscopic variables can also make the observed cellular process non-Markovian; replacing it with a time-homogeneous CTMC requires an approximation to be declared and tested. The strong-lumpability analysis of Section 21 provides an exact finite criterion for one kind of Markovian projection, but does not make such a condition universal in biological systems.

32.4 Scale Consistency Does Not Establish a Condensate

Suppose a family of effective models P_ℓ is specified at spatial or genomic resolutions ℓ and ℓ' , with an explicit coarsening map $C_{\ell \rightarrow \ell'}$ between their state spaces. A directly testable consistency defect is

$$\Delta_{\ell, \ell'} = \sup_{\substack{f \in \mathcal{F} \\ \|f\|_\infty \leq 1}} \left| \mathbb{E}_{(C_{\ell \rightarrow \ell'})_* P_\ell} f - \mathbb{E}_{P_{\ell'}} f \right|, \quad (132)$$

where $\mathcal{F}$ is a declared class of bounded test observables. If $\mathcal{F}$ includes all bounded measurable functions, the expression is twice the usual total-variation distance (under the convention $\text{TV}(P, Q) = \sup_A |P(A) - Q(A)|$). For a restricted set of contact or topology summaries, small Δ proves agreement only for that selected class. Empirical estimation further requires propagating

observational and fitting uncertainty. When no well-defined chain or stochastic map exists across resolutions, the defect is not meaningful until the correspondence is constructed.

Remark 32.2 (Three distinct physical claims). *Agreement between resolutions under Eq. (132) is a distributional consistency statement. Convergence of rescaled Hodge Laplacians or geometrical distances to operators on a limiting space would be a stronger geometric continuum statement, requiring controlled domain, topology, and scaling assumptions. A condensate or phase-transition statement is different again: it requires a defined collective state and suitable gauge-invariant observables with a controlled many-degree-of-freedom limit or other justified phase diagnostic. None follows from the other two without an additional theorem and assumptions. In particular, the independent smooth-field benchmark of Section 9 establishes neither a GFT-origin limit nor a chromatin phase transition.*

An eventual microscopic-to-effective derivation might define an effective action by integrating a nonnegative microscopic statistical weight over fibers of a specified coarse-graining map, schematically

$$e^{-S_\ell^{\text{eff}}(z;e)} := \int_{C_\ell^{-1}(z)} e^{-S_{\text{mic}}(\omega;e)} d\mu_{\ell,z}(\omega), \quad (133)$$

with compatible reference/disintegration measures, a finite integral, and consistent normalization across resolutions. This expression is an *optional definition under assumptions*, not a calculation accomplished in the present paper. The real-time oscillatory weights of a quantum GFT cannot be substituted uncritically into this formula and interpreted as ordinary probabilities. The observable equivalence class of effective descriptions may also depend on the environment, so no scale-independent fundamental group G is assumed by this construction.

32.5 The Predictive Value of Added Topological Descriptors

Whether a more elaborate representation contains useful information is an empirical question with a precise information-theoretic benchmark. Let X be the features made available to a baseline predictor, T a proposed topology-derived descriptor, and Y an independently measured *finite-valued* endpoint on a held-out distribution. Assume that the relevant conditional entropies are finite and evaluation uses ideal expected logarithmic loss over unrestricted conditional predictors.

Proposition 32.3 (Optimal predictive gain and conditional information). *The difference between the optimal expected negative log likelihood using X alone and that using (X, T) is*

$$\inf_{q(\cdot|X)} \mathbb{E}[-\log q(Y | X)] - \inf_{r(\cdot|X,T)} \mathbb{E}[-\log r(Y | X, T)] \quad (134)$$

$$= H(Y | X) - H(Y | X, T) = I(Y; T | X) \geq 0. \quad (135)$$

If T is almost surely a deterministic measurable function of X , then $I(Y; T | X) = 0$.

Proof. For logarithmic loss the true conditional law uniquely minimizes expected loss up to events of zero probability; the optimum is the corresponding conditional entropy. The entropy identity gives the equality and nonnegativity. If $T = f(X)$, conditioning on (X, T) gives the same conditional distribution as conditioning on X , and hence the conditional information is zero. $\square$

This result is consequential for claims of methodological novelty. When a descriptor is calculated deterministically from *exactly the same full input* already available to an unrestricted, ideally trained predictor, it cannot create new Bayes-optimal information. It can nevertheless supply a computationally useful inductive bias, improve data efficiency with restricted predictors, expose interpretable localized features, or reduce cost; each benefit needs its own controlled test. If T incorporates additional independent contact or higher-order measurements not present in X , a positive $I(Y; T | X)$ is possible, but the gain must then be credited partly to that added evidence

and evaluated against a baseline receiving the same evidence in a simpler form. Information gain alone does not imply causal relevance: a perturbation design or other identification assumptions are needed for mechanistic interpretation.

For illustration, if X contains no information about an independent equiprobable binary T and Y equals T with probability 0.85 (otherwise its complement), the ideal logarithmic-score gain is

$$I(Y; T | X) = \log 2 + 0.85 \log 0.85 + 0.15 \log 0.15 \approx 0.27044 \text{ nats.} \quad (136)$$

This example is illustrative, not a measured gain on genomic data. The meaningful biological test is whether a selected, properly controlled cellular descriptor predicts withheld structure or function more reliably or more efficiently than matched contact, graph, sequence and polymer baselines described in Sections 30–31.

32.6 Empirical Discriminators and Research Stopping Conditions

The two directions should converge through quantitative, reproducible discriminators rather than a resemblance of diagrams. Table 5 states four increasing evidential levels. A proposed microscopic claim must pass the tests at its own level; success on a lower level is not transferable as proof of the higher one.

Table 5: Evidence required to advance from constructed genome topology to a physical generative interpretation. These are proposed discriminators, not completed biological results.

Claim level	Minimum evidence	Decisive negative control or failure mode
Cellular validity	Auditable construction rule, oriented incidence, exact $B_1 B_2 = 0$, compatible filtration and replicate sensitivity	Same Betti signature from biologically incompatible face rules; missing or invalid support
Incremental empirical value	Held-out measurements with matched inputs, calibrated uncertainty, localized interpretation and resource cost	Matched contact graph, Potts, persistence, pangenome or polymer baseline performs equally well
History mechanism	Calibrated event-rate model predicts time-resolved and intervened observations across conditions	Distinct mechanisms remain observationally equivalent; phenomenological rates fail transport
Physical condensate origin	Defined microscopic dynamics and gauge-invariant collective/continuum observables with controlled limit and biological mapping	No physical derivation, no distinguishing scaling behavior, or agreement only after post-hoc fitting

The immediate computational project is therefore a measured, data-informed cellular analysis with sparse boundary matrices and comparison against existing tools. Its success criterion is a reproducible and independently useful structural or functional result at a stated resolution. A microscopic GFT derivation forms a separate theoretical criterion. A stronger second project would infer event histories from longitudinal interventions while controlling alignment, topology-construction and measurement uncertainty. Only a third, distinct theoretical project would attempt to derive the cellular law and an emergent geometry from an interacting microscopic model. These efforts can inform each other while retaining separately reportable evidence.

There are also clear stopping rules. The additional combinatorial cost is warranted when topology-conditioned models improve matched benchmarks or yield interpretable invariants robust to data processing. An exactly reducible gauge lift can be eliminated from performance-critical computation while retaining its mathematical interpretation. The title’s stronger physical language remains a hypothesis until the proposed continuum or condensate sector is connected to physical observables and a controlled limit. The manuscript’s mathematical results are conditional constructions; evidence for a fundamental genomic spacetime requires a distinct physical investigation.

33 Identifiability, Physical Interpretation, and Model Limits

The present work establishes an explicit chain-complex representation of a fixed-length history support and a controlled mathematical correspondence among a terminal relational graph, gauge-neutral symbolic readout, exactly integrated contact interactions, and a separately defined continuum scaling test. Deriving nucleotide identities from microscopic-field vacuum selection remains an open problem. The mapping in Eq. (11) is a four-state symbolic encoding; the graph is a prescribed support; and the gauge group $\mathbb{Z}_2$ acts as a modeling redundancy. Molecular chemistry and reconstruction from measured Hi-C or polymer trajectories constitute separate empirical tasks. Proposition 5.2 proves that quotienting the gauge presentation recovers precisely the original four-state terminal law, so additional explanatory power requires independently justified physical couplings. In particular, the finite coherent configurations in Section 11 serve as mathematical collective-state analogues; physical condensates require separate evidence.

A more biologically grounded model would introduce base-specific energies, constraints between complementary strands, polymer degrees of freedom $\mathbf{r}_i \in \mathbb{R}^3$, protein-binding variables, and environment-dependent transition rates. The candidate-contact set would then depend on polymer configuration through experimentally or dynamically specified relations. Marginalizing spatial coordinates could generate effective sequence couplings, but causal mechanisms would require perturbation-based identification beyond correlations and fitted couplings. Data could include reporter assays, contact maps, controlled perturbations, and comparative genomic sequences, with holdout evaluation under interventions.

The GFT-inspired extension poses stronger tasks: (i) derive admissible $\mathcal{K}$ histories from a defined group interaction kernel; (ii) identify physically justified gauge-invariant observables; (iii) find a collective regime with a controlled large-system limit; (iv) establish a geometric continuum description that agrees with measured polymer organization; and (v) demonstrate the relationship of emergent structure to observed DNA states while explicitly testing whether the alphabet can emerge from the underlying model. A state-space or generative precedence relation also depends on specified admissible transitions, environmental conditions, and representational equivalences; the model leaves universal origin and fundamental ordering questions open.

Section 21 adds a controlled path-distribution model and shows that time-linked intervention data can separate some finite mechanisms indistinguishable under baseline observations. This result concerns finite controlled path distributions; reconstructing microscopic cellular histories and establishing a condensate mechanism require additional physical evidence. Section 22 introduces actual edge and face attachment/removal as admissible structural events, with a normalized kinetic path law; the exact three-state example shows that matching observed homology can conceal dynamically distinct complexes. The event rates are phenomenological; connecting them to a microscopic action is a further theoretical task.

The title’s phrase “Gauge Symmetry Breaking” identifies a programmatic hypothesis. The finite model establishes gauge-invariant composite readout, while a physical symmetry-breaking mechanism remains to be derived. The model uses gauge-invariant composites and optional global-state selection. A later version could adopt “Gauge-Invariant State Selection” if biochemical and phase-theoretical evidence does not warrant a stronger breaking mechanism. Correspondingly, “spacetime” presently names generative relational and coarse-grained geometric structures; its relationship to gravitational spacetime is outside the scope of the demonstrated model.

34 Experimental Correspondence and Falsification Criteria

34.1 Relational Observables and Biological Measurement Interfaces

The minimal symbolic model has clear mathematical observables—adjacent-label agreement, contact-conditioned agreement, occupied-contact fraction, and block composition. Physical genome

inference additionally requires experimentally calibrated observables and chemical and structural constraints. A biologically informative extension would define a map from experimentally annotated genomic windows to state variables, incorporate a polymer contact kernel constrained by imaging or chromatin-contact measurements, and introduce base-specific and cell-state-dependent interactions. Existing research on the sequence dependence of genome organization and nonequilibrium polymer physics provides motivation for such measurements (Tamon et al., 2025; Forte et al., 2026). The applicability of the toy model to those mechanisms remains to be established empirically.

An evaluation protocol should separate three sources of information: (i) genomic sequence and local motif identity, (ii) measured or modeled three-dimensional contact organization, and (iii) perturbations that alter binding, accessibility, or spatial folding. Training and test data should be separated by genomic region and, where feasible, biological condition to reduce leakage through homologous or nearby regulatory segments. Synthetic gauge-invariance checks are implementation tests, while chromosome-contact prediction and intervention responses are empirical tests.

34.2 Model Comparisons and Failure Conditions

A meaningful comparison would include a local-only sequence baseline, a nonlocal contact model without gauge variables, the gauge-invariant terminal model, and a more physically parameterized spatial or kinetic extension. The present model’s exact contact marginalization supplies one falsifiable mathematical signature: at fixed x , candidate binary contacts are conditionally independent with the logistic probability in Eq. (16). If real data require cooperative or mutually exclusive contacts, higher-order factors or a different physical model are necessary; the existing expression cannot be retained unchanged.

Similarly, a claimed interacting continuum limit would require a specified sequence of measures with increasing N , fixed rescaling rules, and convergence of selected observables under those rules. The required evidence involves controlled scaling and convergent observables beyond visual embeddings and finite-size order parameters. A claimed physical condensate would further require a collective phase characterized by appropriate gauge-invariant order, correlation structure or response functions, with a clear distinction from imposed discretization and finite-volume sampling. A literal biological interpretation of the title awaits these derivations; at present the title denotes the research program.

34.3 A Minimum Complete End-to-End Derivation

The stronger future theory will be complete only when one specified microscopic action or stochastic kinetic law simultaneously determines (i) the probability of admissible cellular histories, including allowable topology-changing moves; (ii) gauge-invariant terminal-state statistics and a chemically informed readout; and (iii) a controlled collective or continuum description with observables comparable to physical measurements. At present, the first component is supplied by a valid but externally chosen two-complex support and an optional stationary algorithmic history kernel; the second is exact for a four-state toy encoding; and the third is restricted to an uncoupled composition theorem plus an independently chosen smooth-field benchmark. The remaining derivations are identified explicitly, allowing subsequent mathematical and computational work to test each proposed connection.

35 Conclusions and Technical Implications

We constructed and exactly verified an oriented cellular two-complex with a designated genomic boundary, and tested a finite end-to-end terminal generative model whose observable sequence is a gauge-invariant readout of relational degrees of freedom, with nonlocal effective couplings generated by integrating stochastic contacts. Exact identities, gauge invariance, numerical

sampling, continuum discretization, and a coarse-graining obstruction have been verified for the specified examples. The result is a consistent small-scale technical starting point for investigating whether genomic sequence descriptions can be recovered from richer physical generative theories. The computational analysis shows explicitly that the current equilibrium contact model reduces to a Potts-type law: a gauge representation alone creates no predictive advantage. Section 20 further shows that distinct contact-cost and interaction-reward parameters may yield the same complete sequence distribution; calibrated contact observables can distinguish them under the stated model assumptions. This puts a formal limit on sequence-only reconstruction of microscopic mechanisms. A complementary, data-informed genomic analysis requires neither solving a microscopic action nor proving condensation: suitably constructed attributed complexes admit homology, persistent homology, relative topology, cellular/sheaf cohomology, and comparative invariants (Sections 14–29). These outputs depend on the chosen face semantics and measurement conditions; a change in sequence attributes alone leaves the ordinary homology of a fixed complex unchanged. The history-aware cellular extension supplies a finite kinetic model for topology-changing attachments and removals, but its rates are prescribed rather than derived from group-field theory. Section 23 further specifies an exact and particle-based Bayesian reconstruction pathway, validating a finite example without claiming a genome-scale implementation. Section 24 now specifies a typed evidence interface, exact sparse incidence-memory accounting, conditional scaling costs, and a staged benchmark ladder without implying that a large-scale implementation has been released. The proposed AI-learning and quantum-computing extensions are conditional future methods, rather than implemented genome-scale capabilities. Section 32 now provides the explicit bridge between a hypothetical microscopic history law and directly constructed genomic topology, proving an observation-channel information bound and identifying when topological descriptors carry incremental predictive information. Genome-scale feasibility calls for sparse, multiscale inference and empirical comparisons with established genome methods (Sections 12–31). Establishing a genuine condensate, deriving biological nucleotide chemistry, and obtaining validated genome-scale predictions remain open scientific tasks.

License and AI-Assistance Disclosure

Copyright and license. © 2026 Wanhong Huang. Except for any third-party material explicitly credited to its original rights holders, the original text, equations, tables, and author-generated figures in this manuscript are made available under the [Creative Commons Attribution 4.0 International \(CC BY 4.0\)](#) license. Appropriate attribution, a link to the license, and an indication of changes are required for reuse. The license applies to this manuscript and its original author-generated figures; it does not, by itself, announce the public release or licensing of a paper-specific software implementation. Third-party materials retain their respective rights and licenses.

AI-assistance disclosure. OpenAI ChatGPT was used to assist with manuscript drafting and language editing, mathematical consistency review, exploratory Python implementation, and preparation of computational figures and diagnostics. Scientific claims are supported by mathematical arguments and independently specified computational checks; AI suggestions and code provide research assistance. The named author retains responsibility for the research questions, the interpretation of results, all assertions and limitations, and the final manuscript. The numerical procedures and their limitations are described in the manuscript; exploratory code may be requested from the author, without a claim of a released implementation. The physical condensate mechanism remains an explicit research hypothesis.

Code availability, correspondence, and reproducibility

The present public research artifact is the manuscript; a formal paper-specific implementation and genome-scale benchmarking suite are planned for future development. Exploratory author-conducted finite-model Python computations were used for the numerical checks, with algorithmic specifications, seeds, and test conditions described in Sections 11 and E. For project context, subsequent development, and possible future implementation releases, see [The Science Infrastructures](https://github.com/huangwanhong-maker/The-Science-Infrastructures) (<https://github.com/huangwanhong-maker/The-Science-Infrastructures>). For requests concerning source code, numerical details, or collaboration, contact Wanhong Huang at huangwanhong@serendip.ngo. The project repository documents the broader research infrastructure. The reported computations concern finite mathematical examples; cohort-scale benchmarks and experimentally calibrated condensate simulations remain future research milestones.

A Reproducible Cellular Chain Matrices and Their Interpretation

Table 6: Exact topological consistency checks for constructed $C_N \times I_T$. In every row, $\partial_1 \partial_2$ is the zero matrix, and the computed Betti vector is $(1, 1, 0)$.

N	T	$\dim C_0$	$\dim C_1$	$\dim C_2$	χ
3	1	6	9	3	0
6	2	18	30	12	0
8	3	32	56	24	0
12	4	60	108	48	0

The exploratory computational implementation constructs integer matrices D_1 and D_2 in the ordered bases of horizontal edges, vertical edges, and square faces. It computes $D_1 D_2$ exactly and checks the coefficientwise equality $D_2 \mathbf{1} = \Gamma_0 - \Gamma_T$. It reports the cell ranks and resulting homology dimensions using finite-dimensional matrix rank. One should distinguish the statement that a history *can be represented* on a valid CW two-complex from a derivation of the physical dynamics that creates or destroys cells. The present support is a fixed product cylinder; topology-changing operations, branching genomic contours, chemical bond formation, calibrated kinetic rates, and a continuum geometry all require additional work. They are not consequences of the incidence identity alone.

B Independent Smooth-Field Numerical Benchmark

The following plot is a verification of the separate smooth-field discretization and is not a derived continuum limit of the interacting four-state sequence law.

C Soft Four-State Potential: Minimum Calculation

Write $r^2 = \sum_a \Phi_a^2$ and $Q = \sum_{a < b} \Phi_a^2 \Phi_b^2 \geq 0$. For fixed r , Q is minimized by concentrating all magnitude in one coordinate, giving $Q = 0$. The radial potential then becomes $-\mu_0^2 r^2 + \lambda r^4$, minimized at $r = v = \mu_0 / \sqrt{2\lambda}$ with energy $-\mu_0^4 / (4\lambda)$. Without a positivity restriction the directions $\pm e_a$ are distinct and there are eight minima. Constraining all components to be nonnegative or defining the physical field as a positive amplitude selects the four positive branches. This calculation alone says nothing about a positive-temperature phase transition or an interpretation of minima as chemical molecules.

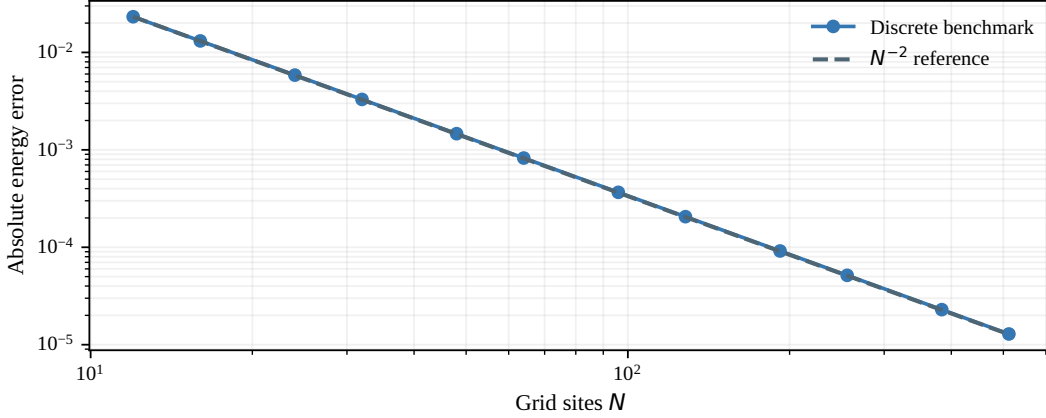


Figure 25: Absolute discrete-to-continuum energy error for the smooth benchmark, with a reference N^{-2} slope. This test confirms a predetermined second-order discretization of a separate smooth-field functional; it does not establish that the four-symbol sequence law has this continuum limit or that a GFT condensate emerges.

D Gauge Orbit Cardinality and Direct Contact Enumeration

For prescribed $q \in \{\pm 1\}^N$, every independent assignment of $h \in \{\pm 1\}^N$ determines the unique links $U_i = h_i q_i h_{i+1}$. Hence there are exactly 2^N microscopic representatives of q . On a finite circle the total product of bond invariants agrees with the link holonomy and imposes no further restriction if holonomy is permitted to vary. Gauge fixing $h_i = 1$ recovers $U_i = q_i$, but the choice of representative is unobservable.

For fixed sequence x , direct enumeration computes

$$\sum_{c \in \{0,1\}^M} \exp \left[-\beta \sum_{e=1}^M c_e (\nu - K \delta_e) \right] = \prod_{e=1}^M \left[1 + e^{\beta(K \delta_e - \nu)} \right],$$

which was independently evaluated through all 2^6 edge choices in the small reference model. The numerical comparison is against independently accumulated floating-point sums, and its reported discrepancy includes floating-point error.

E Algorithmic Specification and Diagnostic Boundaries

1. Initialize x_i independently over four states and all contacts inactive.
2. For each sweep, visit sites in a random permutation. Compute four local conditional weights from Eq. (21), normalize with a maximum-subtraction softmax, and draw one state.
3. For every candidate edge, sample c_{ij} from Eq. (16).
4. After burn-in, retain observables at the declared thinning interval, including adjacent equality, chord equality, occupied-edge fraction, and Q_N .
5. For $N = 6$, enumerate 4^6 strings and use Eq. (17) to calculate exact expectations, then compare with simulation estimates using batch standard errors.
6. Independently test gauge-transform invariance, direct edge elimination, the continuum benchmark, and noncommuting readouts.

The reported test parameters and update scheme are specified in the main text and this appendix. Exploratory implementation details may be requested through huangwanhong@serendip.ngo; readers may also consult the [The Science Infrastructures project repository](#) for broader project context. No publicly maintained paper-specific simulation package is claimed. The tests are computational consistency checks, not proofs of sampler mixing or biological validity. In particular, modest batch errors do not exclude a bias from unvisited metastable regions in substantially larger or lower-temperature models.

References

- Do, C. B., Mahabhashyam, M. S. P., Brudno, M., and Batzoglou, S. (2005). ProbCons: Probabilistic consistency-based multiple sequence alignment. *Genome Research*, 15(2), 330–340. doi:10.1101/gr.2821705.
- Redelings, B. D., and Suchard, M. A. (2005). Joint Bayesian estimation of alignment and phylogeny. *Systematic Biology*, 54(3), 401–418. doi:10.1080/10635150590947041.
- Suchard, M. A., and Redelings, B. D. (2006). BALi-Phy: Simultaneous Bayesian inference of alignment and phylogeny. *Bioinformatics*, 22(16), 2047–2048. doi:10.1093/bioinformatics/btl175.
- Chazal, F., de Silva, V., Glisse, M., and Oudot, S. (2016). *The Structure and Stability of Persistence Modules*. SpringerBriefs in Mathematics. doi:10.1007/978-3-319-42545-0.
- Fletez-Brant, K., Qiu, Y., Gorkin, D. U., Hu, M., and Hansen, K. D. (2024). Removing unwanted variation between samples in Hi-C experiments. *Briefings in Bioinformatics*, 25(3), bbae217. doi:10.1093/bib/bbae217.
- Yardimci, G. G., Ozadam, H., Sauria, M. E. G., et al. (2019). Measuring the reproducibility and quality of Hi-C data. *Genome Biology*, 20, 57. doi:10.1186/s13059-019-1658-7.
- Allman, E. S., Matias, C., and Rhodes, J. A. (2009). Identifiability of parameters in latent structure models with many observed variables. *Annals of Statistics*, 37(6A), 3099–3132. doi:10.1214/09-AOS689.
- Cole, D. J. (2019). Parameter redundancy and identifiability in hidden Markov models. *METRON*, 77, 105–118. doi:10.1007/s40300-019-00156-3.
- Jones, A., Cai, D., Li, D., and Engelhardt, B. E. (2024). Optimizing the design of spatial genomic studies. *Nature Communications*, 15, 4987. doi:10.1038/s41467-024-49174-4.
- Hatcher, A. (2002). *Algebraic Topology*. Cambridge University Press. [Author’s book page](#).
- Kogut, J. B. (1979). An introduction to lattice gauge theory and spin systems. *Reviews of Modern Physics*, 51, 659–713. doi:10.1103/RevModPhys.51.659.
- Mozota Frauca, A. (2024). Foundational issues in group field theory. *Foundations of Physics*, 54, 33. doi:10.1007/s10701-024-00763-9.
- Elitzur, S. (1975). Impossibility of spontaneously breaking local symmetries. *Physical Review D*, 12, 3978–3982. doi:10.1103/PhysRevD.12.3978.
- Forte, G., Brackley, C. A., Gilbert, N., and Marenduzzo, D. (2026). Nonequilibrium polymer models for chromatin. *Current Opinion in Genetics & Development*, 96, 102426. doi:10.1016/j.gde.2025.102426.
- Gielen, S., Oriti, D., and Sindoni, L. (2013). Cosmology from group field theory formalism for quantum gravity. *Physical Review Letters*, 111, 031301. doi:10.1103/PhysRevLett.111.031301.

- Gielen, S., Oriti, D., and Sindoni, L. (2014). Homogeneous cosmologies as group field theory condensates. *Journal of High Energy Physics*, 2014(6), 013. doi:10.1007/JHEP06(2014)013.
- Huang, W. (2026). *Towards a Group Context-Sensitive Grammar for Genome Analysis*. Preliminary technical draft, March 15, 2026, supplied manuscript.
- Tamon, L., Ashford, J., Nicholls, M., de Bruijn, M. F. T. R., and Sahakyan, A. B. (2025). The emerging sequence grammar of 3D genome organisation. *Human Genetics*, 144, 917–928. doi:10.1007/s00439-025-02772-8.
- Michieletto, D., Orlandini, E., and Marenduzzo, D. (2016). Polymer model with epigenetic recoloring reveals a pathway for the de novo establishment and 3D organization of chromatin domains. *Physical Review X*, 6, 041047. doi:10.1103/PhysRevX.6.041047.
- Oriti, D. (2014). Disappearance and emergence of space and time in quantum gravity. *Studies in History and Philosophy of Modern Physics*, 46, 186–199. doi:10.1016/j.shpsb.2013.10.006.
- Kelleher, J., Wong, Y., Wohns, A. W., Fadil, A., Albers, P. K., and McVean, G. (2019). Inferring whole-genome histories in large population datasets. *Nature Genetics*, 51, 1330–1338. doi:10.1038/s41588-019-0483-y.
- Wohns, A. W., Wong, Y., Jeffery, B., et al. (2022). A unified genealogy of modern and ancient genomes. *Science*, 375, eabi8264. doi:10.1126/science.abi8264.
- Salehi Nowbandegani, P., Zhang, S., Hu, H., Li, H., and O’Connor, L. J. (2026). Defining and cataloging variants in pangenome graphs. *Cell Genomics*, 6(9), 101327. doi:10.1016/j.xgen.2026.101327.
- Hickey, G., et al. (2023). Pangenome graph construction from genome alignments with Minigraph-Cactus. *Nature Biotechnology*, 42, 663–673. doi:10.1038/s41587-023-01793-w.
- Di Pierro, M., Zhang, B., Aiden, E. L., Wolynes, P. G., and Onuchic, J. N. (2016). Transferable model for chromosome architecture. *Proceedings of the National Academy of Sciences*, 113, 12168–12173. doi:10.1073/pnas.1613607113.
- OpenMiChroM Developers. (2025). *OpenMiChroM: Chromatin dynamics simulations and documentation*, release 1.1.1. [Project documentation](#).
- Avsec, Z., Latysheva, N., Cheng, J., et al. (2026). Advancing regulatory variant effect prediction with AlphaGenome. *Nature*, 649, 1206–1218. doi:10.1038/s41586-025-10014-0.
- Brix, G., Durrant, M. G., Ku, J., et al. (2026). Genome modelling and design across all domains of life with Evo 2. *Nature*, 652, 1349–1361. doi:10.1038/s41586-026-10176-5.
- Fiorillo, L., Musella, F., Conte, M., et al. (2021). Comparison of the Hi-C, GAM and SPRITE methods using polymer models of chromatin. *Nature Methods*, 18, 482–490. doi:10.1038/s41592-021-01135-1.
- Abdennur, N., and Mirny, L. A. (2020). Cooler: Scalable storage for Hi-C data and other genomically labeled arrays. *Bioinformatics*, 36(1), 311–316. doi:10.1093/bioinformatics/btz540.
- Galitsyna, A. A., Goloborodko, A., Imakaev, M., and Venev, S. V. (2024). Pairtools: From sequencing data to chromosome contacts. *PLOS Computational Biology*, 20(5), e1012164. doi:10.1371/journal.pcbi.1012164.
- Aggarwal, M., and Periwai, V. (2024). Dory: Computation of persistence diagrams up to dimension two for Vietoris–Rips filtrations of large data sets. *Journal of Computational Science*, 79, 102290. doi:10.1016/j.jocs.2024.102290.

- Chiang, Z., and Wu, Y. (2026). Persistent homology-based multiscale Hi-C analysis uncovers TADs-mediated gene regulation. *Scientific Reports*, 16, 22347. doi:10.1038/s41598-026-52335-8.
- Aggarwal, M., and Periwai, V. (2023). Tight basis cycle representatives for persistent homology of large biological data sets. *PLOS Computational Biology*, 19, e1010341. doi:10.1371/journal.pcbi.1010341.
- Cohen-Steiner, D., Edelsbrunner, H., and Harer, J. (2007). Stability of persistence diagrams. *Discrete & Computational Geometry*, 37, 103–120. doi:10.1007/s00454-006-1276-5.
- Carlsson, G., and de Silva, V. (2010). Zigzag persistence. *Foundations of Computational Mathematics*, 10, 367–405. doi:10.1007/s10208-010-9066-0.
- Hansen, J., and Gebhart, T. (2020). Sheaf neural networks. *NeurIPS 2020 Workshop on Topological Data Analysis and Beyond*. arXiv:2012.06333.
- Berry, D. W., Su, Y., Gyurik, C., et al. (2024). Analyzing prospects for quantum advantage in topological data analysis. *PRX Quantum*, 5, 010319. doi:10.1103/PRXQuantum.5.010319.
- McArdle, S., Gilyén, A., and Berta, M. (2026). A streamlined quantum algorithm for topological data analysis with exponentially fewer qubits. *Quantum*, 10, 2058. doi:10.22331/q-2026-04-10-2058.
- Gyurik, C., Schmidhuber, A., King, R., Dunjko, V., and Hayakawa, R. (2026). Provable quantum speedups for computing persistence in topological data analysis. *PRX Quantum*, 7, 020361. doi:10.1103/gvys-hl8h.
- Gjoni, K., Gunsalus, L. M., Kuang, S., McArthur, E., Pittman, M., Capra, J. A., and Pollard, K. S. (2025). Comparing chromatin contact maps at scale: methods and insights. *Nature Methods*, 22, 824–833. doi:10.1038/s41592-025-02630-5.
- Gillespie, D. T. (1977). Exact stochastic simulation of coupled chemical reactions. *Journal of Physical Chemistry*, 81(25), 2340–2361. doi:10.1021/j100540a008.
- Del Moral, P. (2004). *Feynman–Kac Formulae: Genealogical and Interacting Particle Systems with Applications*. Springer. doi:10.1007/978-1-4684-9393-1.
- Doucet, A., and Johansen, A. M. (2011). A tutorial on particle filtering and smoothing: Fifteen years later. In Crisan, D., and Rozovskii, B. (eds.), *The Oxford Handbook of Nonlinear Filtering*, Oxford University Press, pp. 656–704. Author-hosted version.
- Ball, F., and Yeo, G. F. (1993). Lumpability and marginalisability for continuous-time Markov chains. *Journal of Applied Probability*, 30(3), 518–528. doi:10.2307/3214762.
- Park, E.-J., and Kim, H. (2025). Live genome imaging by CRISPR engineering: progress and problems. *Experimental & Molecular Medicine*, 57, 1392–1399. doi:10.1038/s12276-025-01498-x.
- Short, S., García-Tejera, R., Schumacher, L. J., et al. (2025). Next generation lineage tracing and its applications to unravel development. *npj Systems Biology and Applications*, 11, 60. doi:10.1038/s41540-025-00542-w.