Toward a Structural Grammar of Genomes: Learning Regulatory Code with the Phaneron Substrate

Anonymous Preprint

November 13, 2025

Abstract

We propose a single-layer, unlabeled graph substrate—the *Phaneron*—as a general-purpose representation and learning engine for genomic regulation. Nodes carry only distinctness; edges carry only connectivity; all nuance (motifs, n-ary relations, roles, direction, time) is represented by reusable *patterns* (small subgraphs) and *rewrite rules* scored by a hybrid compression–prediction objective. Using **Drosophila** as a first testbed, we target the discovery of regulatory motifs and combinatorial grammar from sequence and ATAC/ChIP profiles, and we evaluate *out-of-tissue* generalization of enhancer→gene linkage. A dynamic pattern-size cap grows abstractions only when they reduce description length and improve held-out prediction. We argue that this yields interpretable, transferable grammars that unify sequence, multi-omics context, and temporal variation within a single structural object.

Contributions.

- A single-layer, unlabeled representation of genomic information in which all semantics live as patterns (no labels/weights on the base graph).
- A hybrid objective combining MDL-style compression with predictive adequacy for ATAC/ChIP and enhancer→gene tasks.
- A rewrite loop that induces motifs and regulatory modules, with a *dynamic pattern-size cap* driven by marginal $\Delta\mathcal{L}$.
- An evaluation protocol emphasizing *out-of-tissue* generalization in **Drosophila**, with strict splits to prevent leakage across tissues and chromosomes.
- A blueprint for scaling to trillion-edge, multi-omic graphs and extending to 3D genome and variant-effect tasks.

1 Introduction

Biological regulation is compositional: short motifs combine into modules, modules combine with 3D context to control genes, and programs vary by cell type, time, and species. We seek a representation that treats "what the genome means" as *structure*, not labels, and that *learns* the grammar by preferring patterns that both compress and predict. The Phaneron offers this: a single-layer unlabeled graph where distinctions (nodes) and connections (edges) are primitive, and all semantics are reusable patterns discovered by a rewrite process.

2 Representation: Genomes as a Phaneron

Distinctions and connections. We model *genome positions* as nodes; sequence adjacency is realized via relation patterns (R_{adjacent}). No labels encode base identities; instead, small patterns (motifs) over positions capture regularities.

Figure 1: **Pipeline.** Sequence and multi-omics are mapped into a single-layer Phaneron; a rewrite loop guided by MDL+prediction induces motifs and modules; we evaluate out-of-tissue (O.O.T.) generalization.

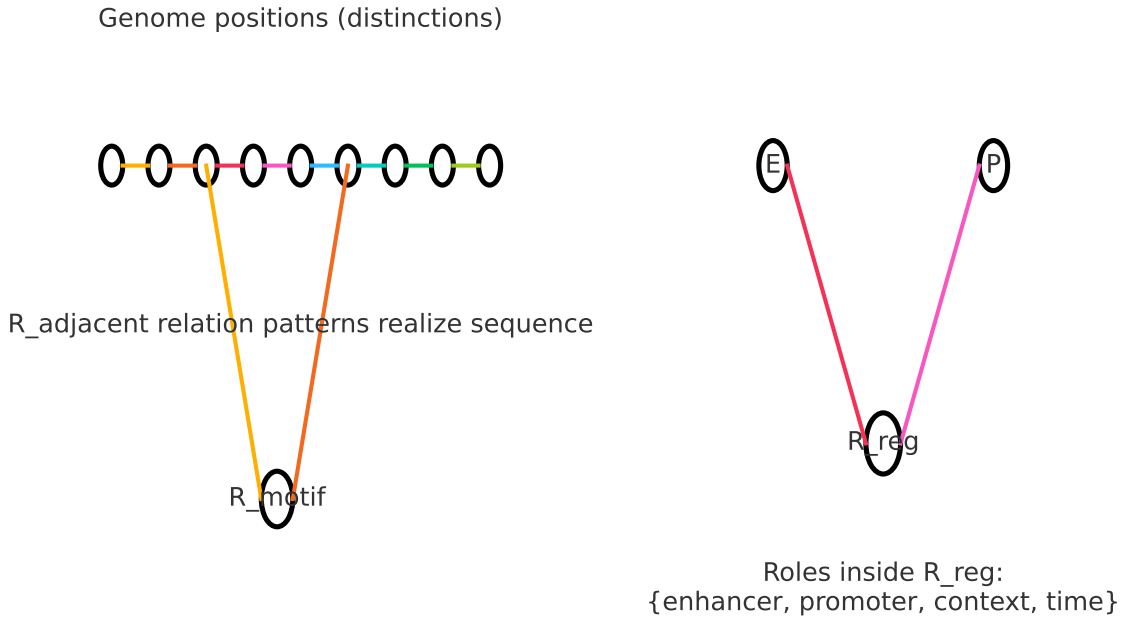

Figure 2: **Single-layer representation.** Positions are distinctions; motifs and regulatory relations are patterns with roles; no base labels or edge arrows leak into the substrate.

All nuance as patterns. Motifs, combinatorial grammars, enhancer↔promoter relations, splicing and translation, 3D loops, and observational context (ATAC/ChIP in tissue c at time t) are all represented as *intermediate relation subgraphs* with role positions. Direction, strand, and temporal order live *inside* these patterns.

Versioning. We treat development, tissue, perturbation, and species axes as versioned events (deltas); replay yields the regulatory "movie" rather than static annotation layers.

3 Objective and Rewrite Semantics

Hybrid objective. We score states and candidate rewrites with

$$\mathcal{L}_{\text{total}} \;=\; \mathcal{L}_{\text{MDL}}(G;\mathcal{D},B) \;+\; \lambda\cdot\mathcal{L}_{\text{pred}}(\hat{Y},Y), \tag{1}$$

where $\mathcal{L}_{\text{MDL}}$ is a two-part code for the pattern dictionary $\mathcal{D}$, optional block structure B, and residuals, and $\mathcal{L}_{\text{pred}}$ measures held-out recovery of structure (masked edges/instances) and multi-omic observations (ATAC/ChIP) under strict splits.

Rewrite loop. Each step mines candidate subgraphs, scores local $\Delta\mathcal{L}$, applies a consistent subset of winners via DPO-style graph rewriting, and appends versioned events; periodic consolidation deduplicates near-duplicate patterns and merges nodes when neighborhoods converge.

Objective:

$$\mathcal{L}_{total} = \mathcal{L}_{MDL}(G;\ \mathcal{D}, B)\ +\ \lambda \cdot \mathcal{L}_{pred}(\hat{Y}, Y)$$

Admissible rewrite S→P if ΔL_total < 0, measured on:

- masked graph structure (static holdout)
- future versions (development/tissue splits)
- tissue O.O.T. prediction (ATAC/ChIP)

Figure 3: **Hybrid objective.** Admissible rewrites reduce total loss and improve held-out recovery across static masks, future versions, and O.O.T. tissues.

Figure 4: **Strict O.O.T. splits.** Train on tissues T1–T4; test on unseen tissues U1–U2; no cross-chromosome leakage.

4 Experimental Design (Drosophila v1)

Data sketch. Genome sequence, ATAC-seq and TF ChIP-seq profiles across multiple tissues / developmental stages (public datasets). Observations are mapped to relation patterns R_{observe} with roles {assay, locus, context, time}.

Tasks. (i) Discover motifs and small regulatory modules that explain ATAC/ChIP; (ii) Predict enhancer→gene linkage in held-out tissues (O.O.T.).

Splits. We partition by *tissue* (train vs unseen test tissues) and *genomic segments* (non-overlapping chromosome blocks) to avoid leakage.

Baselines. Sequence CNN, k-mer logistic/RF, and a graph baseline (sequence graph with labeled edges) trained for the same tasks.

Metrics. AUPRC/AUROC for ATAC/ChIP presence; top-k precision/recall and PR curves for enhancer→gene linkage; calibration for predicted probabilities.

Ablations. MDL-only vs prediction-only vs hybrid; dynamic cap vs fixed cap; positions-as-nodes vs k-mers-as-nodes.

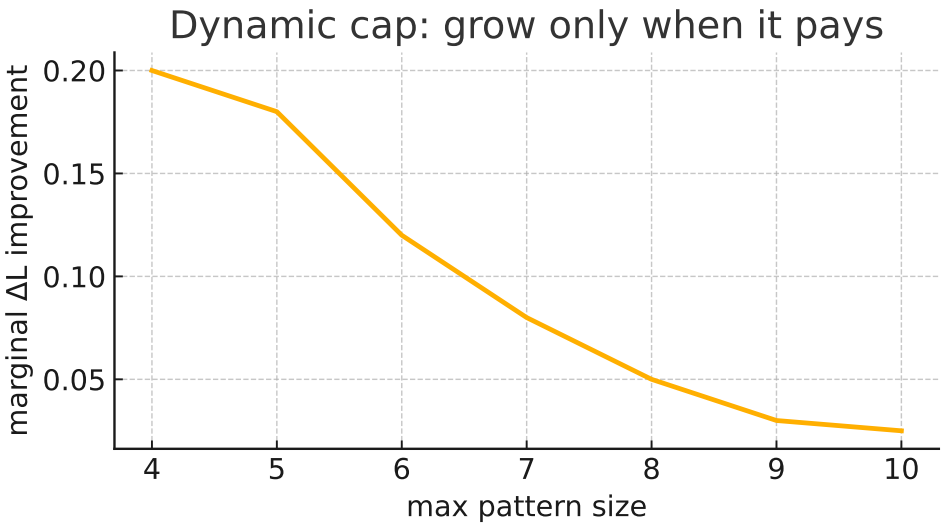

Figure 5: **Grow only when it pays.** Marginal improvement vs maximum pattern size; the cap increases when gains plateau.

5 Dynamic Pattern-Size Cap

We begin with a small cap (e.g., 4–5 nodes) and raise it only when the marginal $\Delta\mathcal{L}$ gain falls below a threshold, ensuring that extra complexity *pays for itself.*

6 Interpretability and Transfer

Every parameter is a concrete pattern with placements; explanations name the subgraph that drives predictions. We expect transfer across tissues (by context-specific placements) and, in v2, across species (by aligning structure rather than surface sequence).

7 Scaling and Systems Notes

Phaneron-Store realizes the ADT with an ID-less, content-addressed state DAG, pattern-first adjacency, reference compression (delta-only), and hot→cold tiers, enabling trillion-edge graphs without label blow-up.

8 Limitations and Ethics

Genomic function depends on chromatin, 3D structure, TF availability, RNAs, and dynamics; sequence-only ablations are not dispositive. MDL can compress artifacts; predictive validation and preregistered splits are required. Variant-effect prediction raises ethical concerns; we recommend a disclosure and safety review for downstream work.

9 Conclusion

A single-layer structural substrate with a hybrid compression–prediction objective provides a unified way to discover regulatory grammar that *generalizes* and *explains*. The **Drosophila** v1 focuses on motifs/grammar with O.O.T. enhancer→gene evaluation; extensions to 3D genome and variant-effect are natural next steps.