Harmonic-Topological Analysis of the Voynich Manuscript as Generator-Class Discrimination: Lossless Glyph Encodings, Graph-Hodge Invariants, and Falsifiable Structural Tests

Derek Earnhart
Independent Researcher
(OriginStoryModel)

April 28, 2026

Complete full-transcription edition integrating the IVTFF/Takahashi EVA experiment, corrected Currier-language vs. hand terminology, and revised BEH results.

Abstract

The Voynich Manuscript remains undeciphered, but its token stream can be analyzed as an unknown symbolic system. This paper presents a lossless harmonic-topological framework for structural analysis and generator-class discrimination of Voynich transcriptions. The framework has three components: (i) an injective phase-amplitude encoding of glyph tokens with exact round-trip recovery; (ii) a graph-Hodge decomposition of token-transition flows whose harmonic core is the well-defined projection onto the cycle space; and (iii) a preregistration-ready falsification protocol that tests structural claims against known ciphers, adversarial pseudo-texts, and cross-corpus controls. We prove that the harmonic core is invariant under monoalphabetic substitution, establishing a hard limitation: it cannot recover plaintext or identify substitution keys. This negative result is central because it separates structural representation from decipherment.

The framework is evaluated on the uploaded IVTFF EVA archive using Takeshi Takahashi's complete `;H` transcription lines. The extraction yields 5,207 Takahashi lines across 225 pages, 37,967 cleaned EVA word tokens, and 8,071 unique word types. Full-transcription results show that Currier B has higher circuit rank and successor entropy than Currier A ($\mu = 13{,}689$ vs. 6,561; $H_{\text{succ}} = 4.199$ vs. 3.382), supporting section- and stratum-level structural discrimination. However, the original one-sided Bimodal Entropy Hypothesis is not supported: Currier A has lower, not higher, page/unit-level entropy than Currier B in this extraction. We therefore revise BEH into a two-sided hand/section entropy-heterogeneity test. Matched null controls show that the actual manuscript has lower circuit rank than frequency-shuffled and uniform-vocabulary controls, indicating order constraints beyond unigram frequency. The framework does not decipher the manuscript. It provides a reproducible measurement layer for excluding implausible generator classes and testing whether the manuscript's structural profile is better explained by natural language, cipher systems, pseudo-text, or procedural symbolic generation. Code and benchmarks are release-ready and should be deposited in a public repository before external review or formal submission.

Contents

Executive Summary

This manuscript presents a complete, updated version of the harmonic-topological Voynich research program. Its central claim is deliberately narrow: the Voynich text can be represented losslessly and compared structurally without claiming decipherment. The framework proves exact round-trip recovery for a phase-amplitude/delay/quaternionic representation, defines a graph-Hodge harmonic core for token-transition flows, proves that this core is invariant under monoalphabetic substitution, and evaluates the method on the full uploaded IVTFF/Takahashi EVA transcription.

The full-transcription experiment changes the manuscript's strongest empirical claim. The original one-sided Bimodal Entropy Hypothesis predicted higher entropy for Currier A or Hand 1; the full extraction does not support that direction. Instead, the evidence supports a revised two-sided heterogeneity claim and a broader generator-class discrimination program: Currier B and Hand 2 show higher transition complexity than Currier A and Hand 1, and the actual manuscript shows lower circuit rank than matched shuffled controls. These findings support structural non-randomness and stratum sensitivity while leaving semantic interpretation unresolved.

1 Introduction

The Voynich Manuscript (hereafter MS 408) is a vellum codex of 209 surviving pages (Beinecke MS 408, Yale University), radiocarbon dated to 1404–1438 [Hodgins, 2011]. It is written in an unknown script using an unknown language—or possibly an elaborate constructed system that mimics a natural language—and has resisted decipherment for over a century despite sustained effort from professional cryptographers, computational linguists, mathematicians, and historians [Beinecke, 2024]. The McCrone Associates materials analysis of the inks and pigments is broadly consistent with a medieval production context [McCrone Associates, 2009].

What *does* the existing literature establish? A great deal, in fact. Computationally, it is known that the Voynich text exhibits lower entropy than random strings [Landini, 2001], that word-frequency distributions follow Zipf-like laws consistent with natural language [Montemurro and Zanette, 2013], that long-range word-order statistics are present [Reddy and Knight, 2011], and that there are at least two distinct scribal hands [Currier, 1976, Davis, 2020]. What has *not* been established is any mapping from Voynich tokens to a natural-language plaintext that commands community consensus.

A recurring failure mode in proposed decipherments is the conflation of two distinct

operations: (a) **encoding**—constructing a reversible, algebraically convenient representation of an unknown symbol stream—and (b) **decoding**—recovering meaningful content from that representation. This distinction has been articulated clearly in the statistical analysis literature: Montemurro and Zanette [Montemurro and Zanette, 2013] demonstrate that information-theoretic methods can certify language-like structure without yielding a single word of plaintext.

The present paper makes three primary methodological contributions and reports one full-transcription application.

1. **A formally specified, invertible representation pipeline.** We define three composable operators on tokenized Voynich text—a complex phase-amplitude encoding, a delay embedding, and a quaternionic rotation— and prove that the composition is lossless (Theorem 3.11).

2. **A graph-Hodge decomposition of token transitions.** We construct the directed token-transition graph, define the edge-flow operator, and apply the Hodge 1-Laplacian to decompose the flow into gradient and harmonic components. Because the transition graph is treated as a one-dimensional complex, no curl component is present. We define the *harmonic core* precisely as the projection onto the nullspace of L_1, characterizing it in terms of the cycle space of the transition graph.

3. **A preregistration-ready falsification protocol.** We specify six ordered tests—from trivially checkable to maximally informative—that any proposed Voynich analysis method must pass before its outputs can be interpreted as evidence of decipherment.

A secondary contribution is a revised *Bimodal Entropy / Heterogeneity Hypothesis* (BEH), which formalizes the long-observed Currier A/B and Hand 1/Hand 2 distinctions [Currier, 1976] as testable claims about differential Shannon entropy. Earlier one-sided framing predicted Currier A and Hand 1 would have higher unit-level entropy; the full IVTFF/Takahashi experiment reported here does not support that direction. The revised form treats entropy separation as a two-sided heterogeneity claim and reserves functional interpretation for later generator-class tests.

We emphasize throughout what the framework cannot do. Hodge decompositions uniquely decompose a given flow; they do not uniquely identify a plaintext. Topological invariants (Betti numbers, cycle ranks) are coarse classifiers; they are not semantic fingerprints. And the phase-amplitude encoding, however elegant, maps unknown symbols to complex numbers without conferring on those numbers any knowledge of what the symbols mean.

Paper organization. Section 2 reviews relevant scholarship on MS 408, distinguishes Currier-language labels from scribal-hand labels, and establishes notation. Section 3 defines the three-stage representation pipeline and proves invertibility. Section 4 develops the graph-Hodge decomposition and the harmonic core. Section 5 states and formalizes the Bimodal Entropy Hypothesis. Section 6 presents the preregistration-ready falsification protocol. Section 9 reports the full IVTFF/Takahashi EVA experiment. Section 10 discusses what is and is not claimed, and Section 11 concludes.

2 Background and Notation

2.1 The Voynich Manuscript and Its Transcriptions

The 209-page codex divides paleographically into several thematic sections: herbal (plant illustrations with surrounding text), astronomical/zodiacal (circular diagrams with radial text), balneological (figures in pools connected by tubes), pharmaceutical (vessel illustrations), and textual/recipe-like pages with marginal star-bullets [Beinecke, 2024]. Folios are not uniformly preserved; the current pagination differs from the original [Zandbergen, 2023].

Computational work depends on transcriptions—machine-readable representations of handwritten glyphs. The European Voynich Alphabet (EVA, Landini 1998) encodes many Voynich signs in ASCII; it remains the most widely used transliteration standard despite known limitations, most notably that it may represent a single glyph as a multi-character group, artificially inflating token counts [Zandbergen, 2023]. The Super Transliteration Alphabet (STA) and the Interlinear Voynich Transliteration File Format (IVTFF) were designed to reduce this ambiguity by providing a richer symbol set that can subsume multiple existing transliterations [Zandbergen, 2023]. Throughout this paper we treat "token" as an abstract unit (a single STA entry or EVA word, depending on the input), and we make explicit which transliteration a given experiment uses.

2.2 Terminology: Currier Language A/B vs. Scribal Hand 1/2

A common source of confusion is the distinction between Currier-language labels and scribal-hand labels. In this paper, **Currier A** and **Currier B** denote statistical/textual strata: differences in token frequencies, preferred word forms, section distribution, and transition behavior. They are not names for physical handwriting styles. By contrast, **Hand 1**, **Hand 2**, and related hand labels denote paleographic or scribal strata: differences in the written forms, ductus, and visual execution of glyphs.

The two systems are correlated because sections written in one textual stratum are often associated with particular scribal hands, but they should not be treated as synonyms. The IVTFF metadata used in Section 9 encodes these separately: `$L` gives Currier language (A/B where known), while `$H` gives Currier hand (1, 2, etc.). For that reason, the empirical tables report Currier A/B and Hand 1/2 separately. The manuscript avoids the ambiguous phrase "Hand A/B" except when discussing earlier informal terminology.

2.3 Key Statistical Results in the Literature

Several empirical facts about MS 408 are well established and guide our framework.

- **Low entropy, high structure.** The per-character entropy of Voynich text, estimated on EVA transcriptions, is below that of random strings and in the range of known natural languages [Landini, 2001].
- **Zipfian word frequency.** Token frequency distributions follow a Zipf-like law, consistent with natural-language word distributions [Montemurro and Zanette, 2013, Reddy and Knight, 2011].
- **Long-range statistical dependencies.** Montemurro and Zanette [Montemurro and Zanette, 2013] apply an information-theoretic method to show that certain words carry long-range statistical dependencies beyond what short-range n-grams capture—a property observed in natural languages but not in random permutations.
- **High daiin frequency.** The token `daiin` (EVA) is among the most frequent tokens and recurs with quasi-periodic regularity in several sections. Its Fourier spectrum shows dominant peaks at specific frequencies, a property we use in Section 5.
- **Currier-language and scribal-hand strata.** Currier [Currier, 1976] identified statistically distinguishable strata in the manuscript. In modern computational use, *Currier A/B* refers primarily to text/statistical-language classes, while *Hand 1/2* refers to scribal or paleographic hand labels. These strata correlate with section divisions but are not identical, so this paper reports them separately.
- **The Naibbe cipher.** Greshko [Greshko, 2025] demonstrates that a medieval cipher called the Naibbe can generate text with Voynich-like statistical properties. This does not decipher the manuscript but establishes that complex structured cipher systems were historically available.

2.4 Notation

Let Σ denote a finite alphabet of K distinct glyph tokens. A *text* is a sequence $\mathbf{g} = (g_1, g_2, \ldots, g_T) \in \Sigma^T$. We write $\mathbb{S}^1 \subset \mathbb{C}$ for the complex unit circle. For a vector $v \in \mathbb{C}^n$ we denote the Hermitian norm $\|v\|^2 = \sum_i |v_i|^2$. The graph Laplacian and Hodge operators are defined in Section 4.

3 The Harmonic Representation Pipeline

The pipeline consists of three composable stages. We specify each stage as a formal operator and prove that their composition is invertible (lossless).

3.1 Stage 1: Phase-Amplitude Encoding

Definition 3.1 (Phase-Amplitude Map Φ)**.** Let $\Sigma = \{s_1, \ldots, s_K\}$ (tokens ordered arbitrarily but fixed). Define unique phases

$$\theta_k = \frac{2\pi(k-1)}{K}, \quad k = 1, \ldots, K,$$

and amplitudes $A_k > 0$, $k = 1, \ldots, K$, with $A_k \neq A_j$ whenever $\theta_k = \theta_j$ (vacuously satisfied here since phases are already distinct). The *phase-amplitude map* is

$$\Phi : \Sigma \to \mathbb{C}, \quad \Phi(s_k) = A_k \, e^{i\theta_k}.$$

For a text $\mathbf{g} \in \Sigma^T$ define the complex wave

$$z_t = \Phi(g_t), \quad t = 1, \ldots, T.$$

Remark 3.2. The map Φ is injective by construction because all pairs (θ_k, A_k) are distinct. The most natural choice is uniform amplitudes $A_k = 1$, reducing Φ to a pure phase encoding on $\mathbb{S}^1$.

Definition 3.3 (Inverse Map Φ^{-1})**.** Given $z \in \mathbb{C}$, the decoder returns

$$\Phi^{-1}(z) = \arg\min_{s_k \in \Sigma} d_{\mathbb{C}}(z, \, \Phi(s_k)),$$

where $d_{\mathbb{C}}(z, w) = |z - w|$ is Euclidean distance in $\mathbb{C}$.

3.2 Stage 2: Delay Embedding

Definition 3.4 (Delay Embedding $\mathcal{E}_{m,\tau}$)**.** For $m \in \mathbb{N}$ (embedding dimension) and $\tau \in \mathbb{N}$ (lag), define

$$\mathcal{E}_{m,\tau} : \mathbb{C}^T \to \mathbb{C}^{N \times m}, \quad N = T - (m-1)\tau,$$

by

$$[\mathcal{E}_{m,\tau}(\mathbf{z})]_{n,j} = z_{n+(j-1)\tau}, \quad n = 1, \ldots, N;\ j = 1, \ldots, m.$$

This construction, attributed to Takens [Takens, 1981] and formalized by Sauer et al. [Sauer et al., 1991], is standard in dynamical-systems analysis of time series. Its relevance here is as a linear reorganization of the data into a higher-dimensional matrix, which allows richer spectral analysis at the cost of reducing the effective sequence length by $(m-1)\tau$.

Remark 3.5. The choice $m = 101$ used in preliminary experiments is motivated operationally (a moderate window large enough to capture long-range token dependencies) rather than by Takens' embedding theorem, which requires $m > 2d$ where d is the dimension of an underlying attractor. No attractor dimensionality is assumed for Voynich text.

3.3 Stage 3: Quaternionic Rotation

Definition 3.6 (Plane-Preserving Quaternion Rotation)**.** Let $\phi \in [0, 2\pi)$ and define the unit quaternion

$$q = \left(\cos\tfrac{\phi}{2},\ 0,\ 0,\ \sin\tfrac{\phi}{2}\right),$$

representing a rotation about the z-axis by angle ϕ. For $c = a + ib \in \mathbb{C}$, define

$$\mathcal{R}_q(c) = e^{i\phi}\, c.$$

Applied element-wise to $X \in \mathbb{C}^{N \times m}$, this defines $\mathcal{R}_q : \mathbb{C}^{N \times m} \to \mathbb{C}^{N \times m}$.

Remark 3.7 (Why Plane-Preserving)**.** We restrict to the plane-preserving subgroup ($q_x = q_y = 0$) for three reasons. First, multiplication by $e^{i\phi}$ keeps every complex number within $\mathbb{C}$, eliminating the need to project back from $\mathbb{R}^3$ and the attendant loss of the z-component. Second, the map is exactly invertible: $\mathcal{R}_q^{-1}(c) = e^{-i\phi}c$. Third, norm preservation is immediate: $|e^{i\phi}c| = |c|$. General quaternion rotations (arbitrary axis) rotate the vector $(a, b, 0)$ to (a', b', c') with $c' \neq 0$ in general, so the "new complex number" $a' + ib'$ discards information in c', violating losslessness.

Lemma 3.8 (Norm Preservation)**.** *For all $c \in \mathbb{C}$ and $\phi \in \mathbb{R}$, $|\mathcal{R}_q(c)| = |c|$.*

Proof. $|e^{i\phi}c| = |e^{i\phi}| \cdot |c| = 1 \cdot |c|$. □

3.4 Invertibility of the Full Pipeline

Definition 3.9 (Coverage Condition)**.** The delay embedding has *full coverage* if

$$N = T - (m-1)\tau \;\geq\; \tau.$$

Under this condition, consecutive embedded windows $[1, N]$, $[1{+}\tau, N{+}\tau]$, ... overlap or touch, so every index $t \in \{1, \ldots, T\}$ appears in at least one embedded coordinate.

Remark 3.10. The coverage condition $N \geq \tau$ is strictly weaker than $T - (m-1)\tau = \tau$ and holds for a wider range of (m, τ, T) triples. When it does not hold, round-trip recovery is still achievable from column $j = 1$ alone; only the recovered length changes from T to N.

Theorem 3.11 (Exact Round-Trip Recovery)**.** *Let* $\mathbf{g} \in \Sigma^T$ *and define* $X = \mathcal{R}_q\big(\mathcal{E}_{m,\tau}(\Phi(\mathbf{g}))\big)$. *Assume:*

(i) $\mathcal{R}_q$ *is plane-preserving (Definition 3.6);*

(ii) the codebook $\Phi(\Sigma)$ *has minimum separation* $\delta = \min_{j \neq k} |\Phi(s_j) - \Phi(s_k)| > 0$*;*

(iii) the final reconstruction error η *(after quaternion inversion and embedding recovery) satisfies* $\eta < \delta/2$.

Then the full decoding pipeline applied to X *recovers* $\hat{g}_t = g_t$ *for all* $t = 1, \ldots, N$*; and for all* $t = 1, \ldots, T$ *if the coverage condition (Definition 3.9) holds.*

Proof. Step 1 (Quaternion inverse). By Definition 3.6, $\mathcal{R}_q(c) = e^{i\phi}c$ for all $c \in \mathbb{C}$. The inverse is $\mathcal{R}_q^{-1}(c) = e^{-i\phi}c = \mathcal{R}_{q^{-1}}(c)$. Since $|e^{-i\phi}| = 1$, applying $\mathcal{R}_{q^{-1}}$ preserves the magnitude of any existing perturbation; all numerical error after inverse rotation and embedding recovery is absorbed into the final reconstruction error η.

Step 2 (Embedding inversion). Column $j = 1$ of $\mathcal{E}_{m,\tau}(\mathbf{z})$ contains $\{z_1, \ldots, z_N\}$, recovering positions 1 through N. Under the coverage condition, combining all m columns recovers $\{z_1, \ldots, z_T\}$.

Step 3 (Nearest-neighbor decoding). Let $\tilde{z}_t$ be the recovered complex number after Steps 1–2. By assumption (iii), the total accumulated error satisfies $|\tilde{z}_t - z_t| \leq \eta < \delta/2$. Since $|\Phi(s_j) - \Phi(s_k)| \geq \delta$ for all $j \neq k$, the ball of radius $\delta/2$ around $\tilde{z}_t$ contains exactly $\Phi(g_t)$, so the nearest-neighbor decoder uniquely recovers g_t. □

Remark 3.12 (Empirical Confirmation)**.** A proof-of-concept experiment encoding `HELLOWORLD` ($T = 10$, $K = 7$ unique tokens) with uniform amplitudes, $m = 5$, $\tau = 1$ achieved 100% round-trip accuracy across 60 trials, consistent with the theorem.

Remark 3.13 (Encoding vs. Decipherment)**.** Theorem 3.11 guarantees that the pipeline is a *lossless representation of the token stream*. It carries no semantic content. Mapping the token `daiin` to the complex number $e^{i\pi/3}$ does not advance our knowledge of what `daiin` means. This distinction is not pedantic; it is the decisive boundary between algebraic signal processing and linguistic decipherment.

Purpose of the Representation. The goal of this pipeline is not compression or encryption, but the construction of a *lossless geometric representation* in which structural invariants of the token sequence can be analyzed independently of the underlying alphabet. In particular, the representation enables the extraction of graph-theoretic and topological features that are invariant under symbol relabeling, making it suitable for the structural analysis of unknown scripts such as the Voynich Manuscript. The value of the pipeline is not that it decodes—it provably cannot—but that it provides a stable, reproducible, parameter-documented coordinate system in which questions about transition structure, cyclic organization, and sectional heterogeneity can be posed and answered without presupposing knowledge of the script's meaning.

4 Graph-Hodge Decomposition and the Harmonic Core

4.1 Token Transition Graph

Definition 4.1 (Transition Graph $\mathcal{G}$)**.** Given text $\mathbf{g} \in \Sigma^T$, define the directed, weighted graph $\mathcal{G} = (V, E, w)$ by:

$$\begin{aligned} V &= \Sigma \quad \text{(unique token types)}, \\ E &= \{(s_i, s_j) \in \Sigma^2 : \exists\, t \text{ s.t. } g_t = s_i,\ g_{t+1} = s_j\}, \\ w(s_i, s_j) &= \text{number of times token } s_j \text{ follows } s_i \text{ in } \mathbf{g}. \end{aligned}$$

This is a standard construction in computational linguistics [Markov, 1913]. For Voynich text, $|V|$ (vocabulary size on EVA transcriptions) ranges from $\approx$ 8,000 for the full manuscript down to a few hundred for individual sections.

Remark 4.2 (Self-Loops)**.** Self-transitions ($s_i \to s_i$), which occur when the same token appears in consecutive positions, are *excluded* from the Hodge incidence graph. In the standard node-edge incidence matrix B_1, a self-loop contributes both head and tail at the same vertex, causing both entries to cancel; this renders the circuit rank ill-defined for graphs with self-loops. Self-transitions are instead stored separately as a *retention statistic*

$r_i = \text{count}(g_t = g_{t+1} = s_i)/\text{count}(g_t = s_i)$, which is an additional structural feature useful for characterizing token "stickiness" across sections.

4.2 Edge Flows and the Hodge 1-Laplacian

Graph Hodge theory [Jiang et al., 2011, Lim, 2015] generalizes the classical Hodge decomposition to simplicial complexes and, in particular, to the 1-skeleton (graph) of a simplicial complex. We work with the undirected version of $\mathcal{G}$ for the Hodge decomposition; the directed flow information is encoded in a signed edge-flow vector.

Definition 4.3 (Incidence Matrix and Edge Flow)**.** Fix an arbitrary orientation of each edge $e \in E$. The *node-edge incidence matrix* $B_1 \in \mathbb{R}^{|V|\times|E|}$ is

$$[B_1]_{v,e} = \begin{cases} +1 & \text{if } v \text{ is the head of } e, \\ -1 & \text{if } v \text{ is the tail of } e, \\ 0 & \text{otherwise.} \end{cases}$$

An *edge flow* is a vector $f \in \mathbb{R}^{|E|}$. The *circulation vector* corresponding to the transition counts is

$$f_e = w(s_i, s_j) - w(s_j, s_i) \quad \text{for edge } e = (s_i, s_j),$$

capturing the net directed flow along each edge.

Definition 4.4 (Hodge 1-Laplacian)**.** The *Hodge 1-Laplacian* (on the graph, without higher simplices) is

$$L_1 = B_1^\top B_1 \in \mathbb{R}^{|E|\times|E|}.$$

Theorem 4.5 (Hodge Decomposition on Graphs, Jiang et al. 2011)**.** *On a graph treated as a 1-dimensional simplicial complex, there are no 2-simplices, so the curl component is absent. Any edge flow $f \in \mathbb{R}^{|E|}$ decomposes orthogonally as*

$$f = B_1^\top \phi + h, \quad h \in \text{Ker}(B_1), \quad B_1^\top \phi \perp h,$$

where $\phi \in \mathbb{R}^{|V|}$ is a node potential ($f_{\text{grad}} = B_1^\top \phi$, the gradient component) and $h = f_{\text{harm}} \in \text{Ker}(B_1)$ is the harmonic component*.*

Remark 4.6. The full three-way Helmholtz decomposition $f = f_{\text{grad}} + f_{\text{curl}} + f_{\text{harm}}$ applies when the complex has triangles and higher faces [Lim, 2015]. On a graph, $f_{\text{curl}} = 0$ identically, and $\text{Ker}(B_1)$ is the cycle space.

Definition 4.7 (Harmonic Core $\mathcal{H}(\mathbf{g})$)**.** The *harmonic core* of text $\mathbf{g}$ is the harmonic component of its circulation flow:

$$\mathcal{H}(\mathbf{g}) = P_{\mathrm{Ker}(L_1)} f,$$

where $P_{\mathrm{Ker}(L_1)}$ is the orthogonal projector onto $\mathrm{Ker}(L_1)$.

Proposition 4.8 (Cycle Space Characterization)**.** $\mathrm{Ker}(L_1) = \mathrm{Ker}(B_1)$ *is precisely the cycle space of the graph* $\mathcal{G}$*. Its dimension equals the* circuit rank *(cyclomatic number)* $\mu = |E| - |V| + c$*, where* c *is the number of connected components.*

Proof. $L_1 f = 0 \iff B_1^\top B_1 f = 0 \iff \|B_1 f\|^2 = 0 \iff B_1 f = 0$, so $\mathrm{Ker}(L_1) = \mathrm{Ker}(B_1)$. By the rank-nullity theorem, $\mathrm{null}(B_1) = |E| - \mathrm{rank}(B_1) = |E| - (|V| - c)$, yielding the cyclomatic number. □

Corollary 4.9 (Structural Interpretation)**.** *The harmonic core* $\mathcal{H}(\mathbf{g})$ *measures the extent to which the transition flow circulates in closed loops that cannot be explained by a node potential. High harmonic energy corresponds to token sequences that tend to revisit cyclic transition patterns; low harmonic energy corresponds to mostly acyclic, "flowing" text.*

Remark 4.10 (What the Harmonic Core Does Not Do)**.** Theorem 4.5 guarantees that $\mathcal{H}(\mathbf{g})$ is a well-defined, reproducible decomposition of the transition flow. It does *not* imply:

(a) that $\mathcal{H}(\mathbf{g})$ uniquely determines a plaintext;

(b) that texts with the same harmonic core are translations of each other;

(c) that texts with different harmonic cores are semantically unrelated.

The decomposition is a structural property of the token-transition dynamics, not a semantic fingerprint. In particular, two texts related by a monoalphabetic substitution cipher will have *identical* harmonic cores up to relabeling of nodes, because substitution preserves transition structure. This is useful as a consistency check but cannot distinguish among valid plaintexts.

4.3 Structural Discriminability of Circuit Rank

The harmonic invariance theorem (Proposition 7.1) establishes what circuit rank *cannot* do. We now establish what it *can*: discriminate between structurally distinct text classes without knowledge of the alphabet or language.

Proposition 4.11 (Circuit Rank Ordering)**.** *For a text* $\mathbf{g} \in \Sigma^T$ *with* $|\Sigma| = K$ *distinct tokens, the circuit rank satisfies*

$$0 \;\le\; \mu(\mathbf{g}) \;\le\; \binom{K}{2} - K + 1 \;=\; \tfrac{K(K-1)}{2} - K + 1.$$

The lower bound is achieved by texts that induce a tree-structured transition graph (no cycles); the upper bound is approached by texts that realize nearly all $K(K-1)/2$ *possible undirected bigrams.*

Proof. $\mu = |E| - |V| + c \geq 0$ (forests have $\mu = 0$). The maximum number of undirected edges on K nodes is $\binom{K}{2}$; a connected graph on K nodes has $c = 1$, so $\mu_{\max} = \binom{K}{2} - K + 1$. □

Corollary 4.12 (Structural Separability)**.** *Texts with strongly Zipfian token distributions (few tokens above the frequency threshold for forming bigrams) have smaller effective vocabulary* $K_{\text{eff}} < K$ *and hence lower* μ *than texts with uniform token frequencies over the same alphabet size.*

This corollary predicts that Voynich-like text (empirically Zipfian, concentrated on ∼30 high-frequency tokens) will have markedly lower circuit rank than uniform random text over the same token inventory—a prediction confirmed by the empirical benchmarks in Section 7.

4.4 Higher-Order Structural Features

Circuit rank μ is the entry-level structural invariant. For finer-grained generator-class discrimination, we define a richer feature family applicable to any token-transition graph.

These features are complementary: circuit rank provides the algebraic foundation via the Hodge decomposition; successor entropy and burstiness provide local and temporal perspectives; spectral peaks test for quasi-periodic structure; and section-conditioned features test the PPH prediction that distinct functional domains exhibit distinct structural profiles (Definition 8.1(ii)).

5 The Bimodal Entropy / Heterogeneity Hypothesis

5.1 Background: Currier A/B and Hand Strata

Currier [Currier, 1976] identified statistically distinguishable textual and scribal strata in MS 408. Subsequent analysis [Davis, 2020] supports a bipartite structure in which Currier A and Currier B differ in glyph morphology, section distribution, and token-frequency profiles.

Table 1: Higher-order structural features for generator-class discrimination.

Feature	Discriminative value
Circuit rank μ	Coarse cyclicity baseline; separates uniform random from Zipfian and natural language
Successor entropy H_{succ}	Mean per-token branching factor; measures local transition predictability
Retention statistic r_i	Self-loop fraction per token; measures "stickiness" without distorting the Hodge graph
Burstiness $B(s)$	$(\sigma - \mu)/(\sigma + \mu)$ over inter-occurrence intervals; $B > 0$ clustered, $B < 0$ periodic
Spectral occurrence peaks	DFT of the binary occurrence vector; identifies quasi-periodic token recurrence
Motif recurrence density	Fraction of length-3 transition motifs appearing ≥ 2 times; detects reused structural templates
Directed role asymmetry	Fraction of tokens with $k_{\text{out}} \gg k_{\text{in}}$; identifies "delimiter-like" or "operator-like" tokens
Section-conditioned μ	Circuit rank computed per section and per hand stratum; tests structural stability vs heterogeneity

In the present paper, we treat these labels as externally supplied strata and ask a narrower statistical question: do those strata differ in unit-level entropy and transition structure?

The original version of the Bimodal Entropy Hypothesis (BEH) made a one-sided prediction: A/Hand 1 units would have higher token-unigram entropy than B/Hand 2 units. The full IVTFF/Takahashi experiment in Section 9 does not support that direction. Therefore the hypothesis is revised here into a two-sided heterogeneity test.

5.2 Formal Statement

Definition 5.1 (Per-Unit Unigram Entropy)**.** For a textual unit U (a page, paragraph-like unit, or other contiguous block in the transcription), let $\hat{p}(s) = \text{count}(s, U)/|U|$ be the empirical unigram frequency of token s. The unit entropy is

$$H(U) = -\sum_{s \in \Sigma} \hat{p}(s) \log_2 \hat{p}(s).$$

Hypothesis 1 (Revised Bimodal Entropy / Heterogeneity Hypothesis (BEH))**.** Let $\mathcal{U}_A$ and $\mathcal{U}_B$ be textual units attributed to Currier A and Currier B, or analogously to Hand 1 and Hand 2, in a public Voynich transcription. The revised BEH asserts a two-sided entropy-separation

claim:

$$\mathbb{E}[H(U) : U \in \mathcal{U}_A] \neq \mathbb{E}[H(U) : U \in \mathcal{U}_B].$$

The direction of the difference is empirical rather than assumed. A functional interpretation of the difference, if any, requires independent validation.

Remark 5.2 (Result of the Original One-Sided BEH)**.** The original one-sided BEH predicted $\mathbb{E}[H_A] > \mathbb{E}[H_B]$. In the full IVTFF/Takahashi run reported in Section 9, that prediction fails at both page and paragraph-like unit levels: Currier A has lower entropy than Currier B, and Hand 1 has lower entropy than Hand 2. This is a negative result for the original directional hypothesis, not a failure of the broader structural-discrimination program.

Remark 5.3 (Motivation and Scope)**.** The revised BEH is a purely *statistical* hypothesis about entropy differences between textual strata. It does not assert that one stratum is a parameter declaration layer, an execution layer, a natural-language variety, or a cipher family. Any such functional interpretation is a downstream Level-3 claim requiring generator-class comparisons and external validation.

5.3 Statistical Test

Definition 5.4 (Two-Sided BEH Test)**.** Let $\bar{H}_A = \frac{1}{|\mathcal{U}_A|} \sum_{U \in \mathcal{U}_A} H(U)$ and $\bar{H}_B = \frac{1}{|\mathcal{U}_B|} \sum_{U \in \mathcal{U}_B} H(U)$. The test statistic is $T = \bar{H}_A - \bar{H}_B$. Under the null hypothesis $H_0 : T = 0$, we reject H_0 at significance level α using a two-sided two-sample permutation test, computing the p-value as the fraction of permutations yielding $|T^*| \geq |T_{\text{observed}}|$.

Remark 5.5 (Falsifiability)**.** The revised BEH is falsifiable: if a two-sided permutation test fails to reject H_0 at a preregistered significance level, or if the observed difference is unstable under reasonable unit definitions and transcription choices, the entropy-heterogeneity claim is not supported. Directional stories about the meaning of the difference are not licensed by the entropy test alone.

5.4 Spectral Structure: DFT Analysis of Frequent Tokens

A separate empirical observation, noted in prior analyses and reproduced here, is that the frequency-domain representation of the positional occurrence vector of high-frequency tokens (such as `daiin`) exhibits prominent spectral peaks—behavior inconsistent with a uniformly random token stream.

Definition 5.6 (Occurrence Spectrum)**.** For token $s \in \Sigma$ and text $\mathbf{g} \in \Sigma^T$, define the binary occurrence vector $\mathbf{x}^{(s)} \in \{0,1\}^T$ by $x_t^{(s)} = \mathbf{1}[g_t = s]$. Its *Discrete Fourier Transform* (DFT) is $\hat{x}_k^{(s)} = \sum_{t=0}^{T-1} x_t^{(s)} e^{-2\pi i k t/T}$. The *occurrence power spectrum* is $S_k^{(s)} = |\hat{x}_k^{(s)}|^2$.

The presence of sharp spectral peaks in $S^{(\texttt{daiin})}$ at specific frequencies k^* would indicate that `daiin` recurs with a characteristic period T/k^*, consistent with either: (a) a regular formatting convention (e.g., a token that marks the end of a structural unit), or (b) a structural periodicity in the underlying content. This observation is consistent with Montemurro and Zanette's [2013] finding of long-range word-order structure.

We do not interpret the identity of `daiin` from this observation; we note only that its spectral structure is a measurable, reproducible property that any proposed theory of the manuscript should account for.

6 Preregistration-Ready Falsification Protocol

The decisive weakness of many proposed Voynich analyses—including earlier versions of our own framework—is the absence of a ground-truth test. Because MS 408 has no accepted plaintext, any method that produces "plausible" output can be mistaken for a decipherment when in fact it is simply a structured transformation that resembles natural language by construction.

We propose a six-tier preregistration-ready falsification ladder. The phrase *preregistration-ready* is essential: criteria must be specified, with quantitative thresholds, *before* the method is applied to Voynich. Post-hoc adjustment of thresholds to match results is not permissible.

6.1 Tier 0: Round-Trip Integrity

Definition 6.1 (T0 Test)**.** Apply the full pipeline to any known text $\mathbf{g}$ and decode. The method passes T0 if and only if token-level reconstruction accuracy is 1.000 (exact).

Status for the present framework: *Pass* (proven by Theorem 3.11; empirically confirmed at 100% on test inputs).

T0 is necessary but trivially insufficient: any information-preserving encoding passes T0.

6.2 Tier 1: Monoalphabetic Cipher Recovery

Definition 6.2 (T1 Test)**.** Let $\mathbf{g}_0$ be a random English text of $\geq 2{,}000$ characters. Apply a random Caesar or monoalphabetic substitution cipher with known key k^* to obtain ciphertext $\mathbf{c}$. Present $\mathbf{c}$ to the method without revealing k^*. The method passes T1 if it recovers k^* (and hence $\mathbf{g}_0$) at accuracy ≥ 0.95 across ≥ 50 independently sampled (g_0, k^*) pairs.

Baseline: Standard frequency-analysis decryption of English monoalphabetic ciphers achieves ≈ 0.98 accuracy on texts of this length, establishing a competitive lower bound.

Status for the present framework: *Evaluated in Section 7.* The Caesar benchmark harness (Algorithm 1) is provided as the reference implementation.

6.3 Tier 2: Polyalphabetic Cipher Recovery

Definition 6.3 (T2 Test)**.** Using a Vigenère cipher with key length $\ell \in \{3, 5, 8, 12\}$, generate ciphertext $\mathbf{c}$ from a known plaintext $\mathbf{g}_0$ (≥ 500 characters per ℓ). The method passes T2 if it recovers the keyword with edit distance ≤ 1 from the true keyword in $\geq 80\%$ of trials.

Status: *Partially evaluated in Section 7.* We evaluate per-position recovery when the key length is assumed known. Full key-length estimation remains unevaluated. The standard approach (index-of-coincidence for key-length estimation followed by per-position Caesar recovery) remains the baseline against which any structural feature should be tested.

6.4 Tier 3: Homophonic Cipher Recovery

Definition 6.4 (T3 Test)**.** A homophonic substitution assigns multiple ciphertext symbols to each plaintext symbol (flattening frequency distributions and mimicking natural language). The method passes T3 if it recovers plaintexts at character accuracy ≥ 0.75 on homophonic ciphertexts of $\geq$ 5,000 characters with known keys.

T3 is specifically relevant because several hypotheses about MS 408 invoke homophonic elements to explain the unusually flat frequency distribution of Voynich tokens [Landini, 2001].

6.5 Tier 4: Adversarial Structured-Noise Discrimination

Definition 6.5 (T4 Test)**.** Generate adversarial pseudo-texts that match Voynich-like statistics (Zipfian word frequencies, low unigram entropy, similar n-gram profiles) but have no underlying plaintext—either by a Markov chain trained on EVA transcriptions, or by the Naibbe cipher mechanism [Greshko, 2025]. The method passes T4 if it correctly classifies $\geq 90\%$ of test samples as "no underlying plaintext" (as opposed to assigning them a specific decipherment) across a balanced set of ≥ 100 pseudo-texts.

Remark 6.6. T4 addresses the most serious confound: a method that "deciphers" Voynich text successfully might simply be finding structure in any structured random text. If the method assigns confident decipherments to adversarial pseudo-texts, its Voynich outputs are unreliable.

6.6 Tier 5: Cross-Corpus Stability

Definition 6.7 (T5 Test)**.** Repeat Tiers 1–4 on non-English plaintext corpora (Latin, Arabic, Medieval Italian) and on synthetic scripts with known statistical properties. The method passes T5 if performance degrades by at most 10 percentage points relative to English baselines.

6.7 Decision Rule

Definition 6.8 (Admissibility Criterion)**.** A method may be presented as *evidence of Voynich decipherment* if and only if it passes Tiers 0 through 3 (T0–T3) with pre-specified thresholds and does not fail T4. Passing T0 alone (as our framework currently does) is *insufficient* for any decipherment claim.

Generator-class use of the ladder. The ladder should not be read as requiring this framework to become a general-purpose decryption system. Tiers 1–3 are control tests: they determine whether a method can recover known plaintexts where ground truth exists and, equally important, whether structural features overclaim within cipher families. The framework's positive use case begins after those controls, where the same features are used to compare generator classes: natural-language text, classical ciphers, homophonic systems, structured pseudo-text, finite-state systems, and candidate procedural grammars.

Algorithm 1 Caesar Cipher Benchmark with Harmonic-Core Feature

Require: Plaintext corpus $\mathcal{C}$, target text length L, trial count $M \geq 50$
Ensure: Key recovery accuracy on 26-key Caesar

Train character trigram LM on $\mathcal{C}$: counts, ctx $\leftarrow$ CharTrigramLM($\mathcal{C}$)
correct $\leftarrow 0$
for $m = 1$ to M **do**
    Sample $\mathbf{g}_0 \sim \mathcal{C}$, truncate to L chars
    Sample key $k^* \sim \text{Uniform}(\{0, \ldots, 25\})$
    $\mathbf{c} \leftarrow \text{CaesarEncrypt}(\mathbf{g}_0, k^*)$
    Build transition graph $\mathcal{G}(\mathbf{c})$ (Definition 4.1)
    Compute $\mathcal{H}(\mathbf{c})$ (Definition 4.7)
    for all $k \in \{0, \ldots, 25\}$ **do**
        $\mathbf{d}_k \leftarrow \text{CaesarDecrypt}(\mathbf{c}, k)$
        $\text{score}_k \leftarrow \text{LMScore}(\mathbf{d}_k) + \lambda \cdot \text{HarmonicConsistency}(\mathcal{H}(\mathbf{c}), \mathcal{H}(\mathbf{d}_k))$
    end for
    $\hat{k} \leftarrow \arg\max_k \text{score}_k$
    if $\hat{k} = k^*$ **then** correct $\leftarrow$ correct $+ 1$
    end if
end forreturn correct$/M$

7 Empirical Evaluation of the Falsification Protocol

We now execute the falsification protocol (Section 6) through Tier 1 and report results honestly. These experiments are not intended to show that the harmonic core improves decryption; they test the *predicted limitations* of the framework. Specifically, the substitution-invariance theorem (Proposition 7.1) predicts zero marginal lift for the harmonic feature on all monoalphabetic tests. The experiments below confirm that prediction in practice and therefore serve as **sanity checks**, validating that the implementation behaves as the theory requires.

7.1 Setup

Corpus. Six passages of English public-domain prose (Melville, Austen, Dickens, Doyle, Shelley, Stoker), totaling approximately 3,000 alphabetic characters, constitute the training corpus for the character trigram language model.

Implementations. Two decoders are compared:

1. **Baseline**: character trigram LM, selecting key $\hat{k} = \arg\max_k \text{LMScore}(\text{Dec}(c, k))$.
2. **Harmonic**: trigram LM augmented with a harmonic-consistency term: $\text{score}_k = \text{LMScore}(\text{Dec}(c, k)) + \lambda \cdot (-|\mathcal{H}(\text{Dec}(c, k)) - \bar{\mathcal{H}}_{\text{Eng}}|)$, where $\bar{\mathcal{H}}_{\text{Eng}} = 0.9997$ is the mean harmonic ratio measured on the English corpus and $\lambda = 0.30$.

Trials. 60 independent trials, each with a freshly sampled 400-character plaintext segment and a uniformly random Caesar key $k^* \in \{0, \ldots, 25\}$.

7.2 Tier 0: Round-Trip Integrity

Result: PASS. Exact token reconstruction (100% accuracy) was achieved on all test inputs, consistent with Theorem 3.11.

7.3 Tier 1: Caesar (Monoalphabetic) Recovery

Both methods pass Tier 1 at ceiling accuracy. The harmonic feature provides **zero marginal lift** on all 60 trials.

Table 2: Tier-1 Caesar recovery results (60 trials, 400-char ciphertext).

Method	Correct	Accuracy	T1 threshold	Verdict
Baseline (LM only)	60/60	1.000	≥ 0.95	**PASS**
Harmonic (LM + $\mathcal{H}$)	60/60	1.000	≥ 0.95	**PASS**
Harmonic lift (Δ)	0/60	0.000	—	(neutral)

Why the harmonic feature is neutral on monoalphabetic ciphers. This result is not a surprise: it is the empirical confirmation of Remark 4.10. A Caesar cipher (and more generally any monoalphabetic substitution cipher) relabels the nodes of the token-transition graph while preserving all edge structure. It is therefore a graph isomorphism. Formally:

Proposition 7.1 (Harmonic Invariance Under Monoalphabetic Substitution)**.** *Let $\sigma : \Sigma \to \Sigma$ be a bijection and $\sigma(\mathbf{g}) = (\sigma(g_1), \ldots, \sigma(g_T))$. Then:*

(i) The transition graphs $\mathcal{G}(\mathbf{g})$ and $\mathcal{G}(\sigma(\mathbf{g}))$ are isomorphic via the node bijection induced by σ.

(ii) The circuit rank $\mu(\sigma(\mathbf{g})) = \mu(\mathbf{g})$.

(iii) $\|\mathcal{H}(\sigma(\mathbf{g}))\| = \|\mathcal{H}(\mathbf{g})\|$.

Proof. The transition count satisfies $w_{\sigma(\mathbf{g})}(\sigma(s_i), \sigma(s_j)) = w_{\mathbf{g}}(s_i, s_j)$ for all $s_i, s_j \in \Sigma$. Therefore $\mathcal{G}(\sigma(\mathbf{g}))$ and $\mathcal{G}(\mathbf{g})$ are isomorphic graphs. Graph isomorphisms preserve $|V|$, $|E|$, and connected components c, hence $\mu = |E| - |V| + c$ is preserved, establishing (i) and (ii). Isomorphism also preserves the cycle space and its dimension, so the harmonic projection norm is preserved, establishing (iii). □

Remark 7.2 (Hard Limitation and Scope)**.** Proposition 7.1 establishes a hard boundary on harmonic-core features: they are *incapable* of distinguishing between plaintext and any monoalphabetic ciphertext thereof, nor of identifying which of the $K!$ relabelings corresponds to the correct key. Consequently, circuit-rank and harmonic-norm features are unsuitable for direct key recovery against substitution ciphers.

This is not a failure of implementation. It is a structural theorem that *precisely delimits valid use cases*: these features are appropriate for cross-class structural analysis (e.g., distinguishing language families, cipher families, or procedural text from linguistic text) but not for within-class key discrimination. Reporting this limitation explicitly prevents a category of reviewer objections and distinguishes this framework from proposals that overclaim.

Proposition 7.1 establishes a hard bound on the harmonic-core feature: it **cannot** help recover a Caesar or substitution cipher key, because the feature value is identical for all $K!$ substitution variants of a given text. The Baseline (frequency analysis via the trigram LM) is the correct and sufficient tool for monoalphabetic ciphers.

7.4 Structural Discriminability of the Circuit Rank

Although the harmonic ratio scalar is uninformative for key selection within a cipher family, the circuit rank $\mu = |E| - |V| + c$ of the transition graph *is* a useful classifier across text types. Table 3 reports measurements on four text categories for a fixed 600-character window.

Table 3: Transition-graph structural features across text types (600-char samples; repeated trial mean for random/proxy).

Text type	H(unigram)	Circuit rank μ	Null-dim(L_1)
English plaintext	2.916	134	134
Caesar ciphertext ($k = 3, 13$)	2.916	134	134
Uniform random text	3.232	256	256
Voynich-statistics proxy	2.495	95	95

Three observations follow directly.

1. **Monoalphabetic invariance is confirmed numerically.** Caesar ciphertexts (both $k = 3$ and $k = 13$) have identical circuit rank to their plaintexts, as predicted by Proposition 7.1.

2. **Circuit rank discriminates text types.** English ($\mu = 134$), Voynich-proxy ($\mu = 95$), and random ($\mu = 256$) are well separated. The Voynich-proxy's lower circuit rank reflects its more constrained, Zipfian token distribution—a smaller fraction of possible bigrams are observed, yielding fewer independent cycles.

3. **Unigram entropy and circuit rank are complementary.** English and the Voynich proxy have similar μ separation but different entropy profiles; random text is the outlier on μ but not dramatically so on entropy. Neither statistic alone is a complete fingerprint; the pair (μ, H) provides richer discrimination.

7.5 Vigenère (Polyalphabetic) Results

Table 4 reports key-recovery accuracy and character-level plaintext accuracy for the baseline and harmonic decoders on Vigenère ciphertexts of key lengths $\ell \in \{3, 5, 8\}$ (40 trials each,

500-char plaintext). The harmonic-consistency term uses circuit rank (rather than the scalar harmonic ratio) as the structural feature.

Table 4: Vigenère recovery results (40 trials per key length, 500-char ciphertext; key-length assumed known for the per-slice recovery step).

Method	Key len ℓ	Key accuracy	Char accuracy
Baseline	3	1.000	1.000
Harmonic	3	1.000	1.000
Δ	3	0.000	0.000
Baseline	5	0.850	0.955
Harmonic	5	0.850	0.955
Δ	5	0.000	0.000
Baseline	8	0.400	0.906
Harmonic	8	0.400	0.906
Δ	8	0.000	0.000

Interpretation. As expected, baseline per-slice recovery degrades with key length: at $\ell = 8$, each slice contains only $500/8 \approx 63$ characters—too short for reliable trigram frequency discrimination across all 26 keys. The harmonic feature provides no lift at any key length tested.

The reason is the same as for Caesar: within each slice, the per-slice decryption step applies a monoalphabetic substitution, and Proposition 7.1 applies to each slice independently. The harmonic-consistency reward cannot distinguish the correct key for a given slice position from an incorrect one with the same structural signature.

What would help. The circuit rank is identical for all monoalphabetic variants of a fixed text, but the *ratio of circuit ranks across positions* in a Vigenère slice decomposition may differ between correct and incorrect key-length hypotheses. This is a direction for future work: using μ not per-slice but as a *consistency check across decomposition lengths* during the key-length estimation step (before per-position recovery).

7.6 Summary: What the Experiments Establish and Refute

1. **T0 (invertibility):** PASS. Proven analytically and confirmed empirically.
2. **T1 (monoalphabetic recovery):** PASS at ceiling (100%), but the harmonic feature is mathematically proven to be uninformative for this tier. The Baseline trigram LM is the appropriate and sufficient tool.

3. **Harmonic-core feature for key recovery:** NEGATIVE RESULT. The feature provides zero marginal lift on both Caesar and Vigenère cipher recovery. This is a consequence of Proposition 7.1, not a failure of implementation.

4. **Circuit rank as a structural classifier:** POSITIVE RESULT. Circuit rank discriminates text types (English, Voynich-proxy, random) without knowing the alphabet or language. This supports the use of μ as a language-agnostic structural invariant in Voynich section analysis.

5. **T2–T4:** T2 is partially evaluated only under the simplifying assumption that key length is known; full key-length estimation remains unevaluated. T3 and T4 remain unevaluated. The Vigenère results at $\ell = 8$ show that key-length estimation is the bottleneck for the Baseline; future work should test whether circuit-rank-based key-length estimation improves T2 performance.

8 From Structural Representation to Procedural-Hypothesis Testing

8.1 Scope Clarification

The preceding sections establish a **lossless symbolic-to-geometric encoding pipeline** and a corresponding **topological analysis framework**. These results are strictly structural in nature:

- the encoding is provably invertible under defined constraints (Theorem 3.11);
- the harmonic/Hodge decomposition yields reproducible invariants of the token-transition structure (Theorem 4.5, Proposition 7.1);
- these invariants are *representation-level properties*, not semantic assignments (Remark 3.13).

  No claim of linguistic decipherment or semantic recovery is made in this paper.

A reversible encoding of a symbol stream into a geometric or algebraic space does not imply that the underlying system has been interpreted. The empirical benchmarks of Section 7 confirm this boundary precisely: the harmonic core is invariant under monoalphabetic substitution, so it cannot recover a substitution key, yet it does discriminate structurally distinct text types via circuit rank.

8.2 Candidate Model Class: The Pasigraphic Procedural Hypothesis

Separately from the structural results above, a broader interpretive framework—developed in companion materials and motivated by paleographic analysis—proposes that MS 408 may not encode a natural language at all, but instead constitutes a *non-phonetic procedural system*. We formalize this as a named hypothesis.

Definition 8.1 (Pasigraphic Procedural Hypothesis (PPH))**.** The Voynich Manuscript is a structured symbolic system in which glyph sequences encode *procedural relationships*—such as ingredients, transformation operations, timing, and constraints—rather than phonetic or lexical content of a natural language. Under the PPH:

(i) glyphs function as conceptual or operational units, not letters;

(ii) distinct manuscript sections correspond to functionally distinct domains (e.g., material specification vs. transformation procedures);

(iii) illustrations act as mnemonic or schematic augmentations of the symbolic system rather than as botanical or astronomical representations in the conventional sense.

The PPH is consistent with several independently documented observations: statistical heterogeneity across Currier-language and scribal-hand strata [Currier, 1976, Davis, 2020]; repetitive, low-entropy token structures [Landini, 2001]; and section-dependent variation in glyph usage [Montemurro and Zanette, 2013].

However:

The PPH is a hypothesis, not a result of the present framework.

The distinction matters. The structural invariants computed here are consistent with the PPH, but they are also consistent with several other hypotheses (natural language cipher, constructed language, meaningless-but-structured pseudo-text). Consistency is not confirmation.

8.3 Role of the Harmonic Framework as a Testing Layer

The contribution of this work is to convert the PPH from a qualitative narrative into a *quantitatively testable model class*. The harmonic-topological framework enables four specific tests.

(i) Sectional discrimination. By analyzing transition graphs and their Hodge decompositions across manuscript sections, we can test whether different sections occupy distinct regions in the structural feature space $(\mu, H, \|\mathcal{H}\|)$. The circuit rank measurements in Table 3 show that a Voynich-statistics proxy is already separated from both English and random text. Applying the same measurement to the actual EVA transcription, stratified by section and hand label, constitutes a direct test of part (ii) of Definition 8.1.

(ii) Structural consistency testing. If the manuscript encodes a procedural system, one expects repeated motifs corresponding to operations or states, and constrained transition structures across contextually similar passages. These expectations are evaluable via circuit rank, per-paragraph entropy (as in the BEH, Section 5), and harmonic component distributions.

(iii) Adversarial falsification. The Tier 4 test (Section 6) requires the framework to classify adversarial pseudo-texts—Markov-generated or Naibbe-cipher outputs [Greshko, 2025]—as lacking an underlying plaintext. If the structural features of MS 408 are indistinguishable from those of adversarial pseudo-texts, the PPH's distinctiveness claim is undermined.

(iv) Invariance analysis. Because monoalphabetic substitution preserves transition structure (Proposition 7.1), the framework identifies precisely what *cannot* be inferred from structural invariants alone. This prevents overinterpretation and constrains the valid hypothesis space: any claim that must appeal to specific glyph identities (Level 3 in Section 8.5 below) cannot be supported by structural evidence alone.

8.4 The Bridging Principle

We can now state the core synthesis of the paper:

> **The harmonic-topological framework provides the formal testing layer for evaluating whether MS 408 behaves like a pasigraphic procedural system—without presupposing that interpretation.**

The PPH generates predictions; the present framework provides tools to test those predictions; only successful, repeatable validation justifies strengthening the hypothesis.

Table 5: Claim hierarchy for the Voynich harmonic-topological framework.

Lvl	Label	Content	Status
0	Established	Lossless encoding; graph-Hodge decomposition; substitution invariance.	Proven (Thms. 3.11,4.5; Prop. 7.1)
1	Supported	Circuit rank and entropy discriminate broad text classes (English, Zipfian, random).	Confirmed by synthetic/proxy benchmarks (Tables 2,3)
2	Full-transcription result	Voynich strata exhibit structurally distinct profiles; original one-sided BEH direction fails and is revised to two-sided heterogeneity.	Tested on IVTFF/Takahashi EVA extraction (Tables 6–8)
3	Hypothetical (PPH)	Non-phonetic procedural system; functionally distinct domains.	Definition 8.1; consistent with evidence; not confirmed
4	Unverified	Specific glyph-to-concept assignments; identification as a historical practice.	Requires external validation; not supported by structural evidence

8.5 Claim Hierarchy

To maintain scientific clarity, we distinguish five levels of assertion.

The methodological advantage of this hierarchy is that it allows interpretively rich hypotheses (including the "paper computer" framing of companion materials) to be explored *without sacrificing scientific rigor*, by grounding them in falsifiable structural tests at Levels 0 and 1.

Representation is solved. Interpretation remains conditional.

9 Full IVTFF/Takahashi EVA Experiment

9.1 Setup and Extraction

The proxy-only analysis was replaced with a full run on the uploaded IVTFF EVA archive, using only Takeshi Takahashi `;H` transcription lines. The archive identifies `H` as Takahashi's complete transcription and provides page-level metadata for illustration type (I), Currier language (L), and Currier hand (H). After removing editorial comments and uncertain filler material, the extraction produced 5,207 `;H` lines across 225 pages, yielding 37,967 cleaned EVA word tokens and 8,071 unique EVA word types. The most frequent token is `daiin`, with 864 occurrences.

Self-transitions are excluded from the Hodge incidence graph and reported separately as retention statistics, as specified in Remark 4.2. Features are computed at full-manuscript, Currier-language, hand, page, and paragraph-like unit levels.

9.2 Aggregate Structural Features

Table 6: Full IVTFF/Takahashi EVA aggregate structural features.

Stratum	Tokens	Vocab	H	μ	H_{succ}	`daiin`
Full manuscript	37,967	8,071	10.452	22,675	4.361	864
Currier A	11,450	3,410	9.865	6,561	3.382	512
Currier B	23,224	4,926	9.886	13,689	4.199	315
Hand 1	7,273	2,213	9.333	4,062	3.245	384
Hand 2	9,704	2,283	9.086	5,493	3.762	148

The aggregate result supports the broader structural-discrimination program. Currier B has substantially higher circuit rank and successor entropy than Currier A ($\mu = 13{,}689$ vs. 6,561; $H_{\text{succ}} = 4.199$ vs. 3.382). Hand 2 likewise has higher transition branching than Hand 1 ($H_{\text{succ}} = 3.762$ vs. 3.245). These results indicate that manuscript strata differ not only in token frequencies but in transition-graph structure.

9.3 BEH Result: Directional Prediction Not Supported

The full-transcription experiment does not support the original one-sided BEH direction. At both page and paragraph-like unit levels, Currier A and Hand 1 have lower mean entropy than Currier B and Hand 2.

Table 7: Full IVTFF/Takahashi entropy tests. The one-sided test shown is the original BEH direction A/1 > B/2.

Test	Mean A/1	Mean B/2	ΔH	One-sided p for A/1 > B/2
Page: Currier A > B	6.016	6.958	-0.943	1.0000
Paragraph-unit: Currier A > B	5.712	6.453	-0.741	1.0000
Page: Hand 1 > 2	5.840	6.590	-0.750	1.0000
Paragraph-unit: Hand 1 > 2	5.726	6.115	-0.389	1.0000

This negative result is scientifically useful. It rules out the original directional version of BEH for this extraction and requires the manuscript to treat entropy as a two-sided heterogeneity measure. The stronger claim that a specific entropy direction corresponds to a functional role, such as "parameter declaration" versus "procedural execution," is not supported.

9.4 Matched Null Controls

Two matched controls were computed. The frequency-shuffle control preserves the exact unigram counts of the manuscript while randomizing order; the uniform-vocabulary control preserves length and vocabulary but samples uniformly.

Table 8: Matched controls for full IVTFF/Takahashi extraction. Values are means over 20 control trials where applicable.

Control	H	μ	H_{succ}	Retention
Actual manuscript	10.452	22,675	4.361	0.838%
Frequency shuffle	10.452	24,506	4.526	0.311%
Uniform vocabulary	12.818	29,941	2.394	0.013%

The actual manuscript has lower circuit rank than the frequency-shuffled control despite identical unigram counts. This supports the claim that observed transition structure is constrained by token order, not merely by vocabulary size or unigram frequency. The uniform-vocabulary control produces much higher entropy and circuit rank, as expected.

9.5 Frequent-Token Recurrence: `daiin`

The token `daiin` occurs 864 times in the cleaned Takahashi extraction. Its full-manuscript inter-occurrence burstiness is $B = 0.214$, indicating clustered rather than perfectly periodic recurrence. Currier A contains more `daiin` tokens than Currier B (512 vs. 315), while B

still exhibits higher overall transition complexity. Thus `daiin` is a useful stratum-sensitive recurrence marker, but its recurrence does not by itself identify semantic content.

Spectral occurrence peaks were computed as diagnostics. Because low-frequency DFT peaks can reflect page ordering, section boundaries, or sampling artifacts, they are not interpreted as semantic periodicities. Their proper use is comparative: the same peak-strength and burstiness measures should be evaluated against matched shuffles, section-stratified permutations, and generator-class controls.

9.6 Implications for the Paper

The full IVTFF/Takahashi experiment changes the manuscript's empirical posture in three ways. First, it upgrades the work from proxy-only diagnostics to a real full-transcription analysis. Second, it refutes the original one-sided BEH direction and motivates the revised two-sided BEH in Section 5. Third, it supports the broader thesis that graph-based structural features can discriminate manuscript strata and detect order constraints beyond unigram frequency.

10 Discussion

10.1 What the Framework Establishes

Section 8 formalizes the complete claim hierarchy. We summarize the Level-0 (established) results here for reference.

Invertibility. The three-stage pipeline—phase-amplitude encoding, delay embedding, quaternionic rotation—is provably lossless (Theorem 3.11), empirically confirmed at T0.

Well-defined harmonic core. The graph-Hodge decomposition (Theorem 4.5) yields a formally defined harmonic component with a clear algebraic interpretation as the cycle-space projection of the transition flow (Corollary 4.9).

Harmonic invariance under substitution. Proposition 7.1, confirmed empirically (Table 2 and Table 3), establishes that the harmonic-core scalar cannot distinguish monoalphabetic cipher variants. This is a structural theorem, not a limitation of implementation: it precisely delimits where harmonic features are and are not informative, and prevents a category of overinterpretation.

Circuit rank as cross-type structural discriminator. Table 3 demonstrates that μ separates English, Voynich-proxy, and random text on matched samples—a language-agnostic invariant with direct applicability to MS 408 section analysis.

Bimodal Entropy / Heterogeneity Hypothesis. The full IVTFF/Takahashi experiment refutes the original one-sided BEH direction and supports revising BEH into a two-sided entropy-heterogeneity test. Entropy differences are real structural observations only after stratified testing; they do not by themselves imply functional roles or semantic content.

Pasigraphic Procedural Hypothesis (PPH). Definition 8.1 formally introduces the PPH as a candidate model class. The full-transcription results show stratum-sensitive structural differences, but they do not confirm the PPH. The PPH remains a generator-class hypothesis to be tested against natural-language, cipher, pseudo-text, and procedural grammar baselines.

10.2 What the Framework Does Not Establish

The following claims are explicitly *not* supported.

Unique translation. The harmonic core is a property of the generative mechanism, not the plaintext. Proposition 7.1 shows it is invariant under the entire class of monoalphabetic substitutions; it cannot certify which (if any) substitution maps to a natural language.

Semantic content from topology. Cycle ranks and Betti numbers are coarse invariants: many non-isomorphic graphs share the same values. Matching these invariants between Voynich and a candidate plaintext is a consistency check, not a proof of decipherment.

Hodge diamond as feature dimension. The parameter $m = 101$ in the delay embedding is a window length, not a Hodge number. Conflating the two misuses mathematical terminology and is disallowed.

Symplectic integration without a Hamiltonian. Symplectic integrators are well-defined for Hamiltonian systems [Leimkuhler and Reich, 2004], but require an explicit $H(q, p)$. Invoking symplectic language without specifying the Hamiltonian names a family, not a computation.

Level-3 claims in general. Specific glyph-to-concept assignments, reconstruction of procedures, and identification as a specific historical practice all require external validation that structural invariants cannot provide (Table 5).

10.3 Position in the Literature

The framework sits at the intersection of two traditions.

The *statistical Voynich* tradition [Montemurro and Zanette, 2013, Reddy and Knight, 2011, Landini, 2001] characterizes structural properties without claiming decipherment. Our contribution is the formally specified harmonic-core invariant, the BEH, and the circuit-rank discriminator—each a reproducible, parameter-light addition to this tradition.

The *graph signal processing* tradition [Defferrard et al., 2016, Barbarossa et al., 2020] applies spectral and topological operators to network data. Our contribution is the first application of the graph Hodge 1-Laplacian [Jiang et al., 2011, Lim, 2015] to the token-transition graph of an unknown script, with a proof of its invariance properties under symbol substitution.

10.4 Path Forward

The immediate experimental priority is no longer simply obtaining a real EVA transcription: that step is now complete for the uploaded IVTFF/Takahashi extraction (Section 9). The next priority is to strengthen the control suite: matched word-level natural-language corpora, Latin and medieval-language baselines, homophonic cipher controls, Markov and hidden-Markov pseudo-texts, and synthetic procedural grammars.

A second priority is robustness analysis. The IVTFF results should be repeated under alternative transcription choices, glyph-level tokenization, word-level tokenization, page-level and section-level partitions, and exclusion/inclusion rules for labels and uncertain readings. Only findings that remain stable across these choices should be treated as manuscript-level structural facts.

Only after those controls are in hand does it become scientifically appropriate to ask whether the structural fingerprint of MS 408 is better described by the PPH than by natural-language, cipher-text, or pseudo-text null hypotheses.

11 Conclusion

We have presented a mathematically explicit, empirically tested, and fully falsifiable analysis framework for the Voynich Manuscript (MS 408). The final framework comprises five integrated contributions.

(1) A provably invertible representation pipeline. Phase-amplitude encoding, delay embedding, and quaternionic rotation compose into a lossless transform (Theorem 3.11), confirmed at T0 (100% round-trip accuracy).

(2) A graph-Hodge harmonic core. The Hodge 1-Laplacian decomposition of the token-transition graph defines the harmonic core as the cycle-space projection of the transition flow (Theorems 4.5, 7.1). Empirical benchmarks establish that this invariant is: (a) identical for all monoalphabetic substitution variants of a text (proven); (b) discriminative across structurally distinct text types via circuit rank (measured; Table 3).

(3) A preregistration-ready six-tier falsification protocol. Tier-0 and Tier-1 tests

are executed and reported honestly, including the negative result that the harmonic-core scalar provides zero marginal lift on Caesar cipher recovery—a consequence of the substitution invariance theorem, not an implementation failure.

(4) The revised Bimodal Entropy / Heterogeneity Hypothesis. The original one-sided BEH prediction is tested on the full IVTFF/Takahashi extraction and is not supported. The manuscript therefore revises BEH into a two-sided hand/section entropy-heterogeneity test, independent of any content interpretation.

(5) The Pasigraphic Procedural Hypothesis as a testable model class. Definition 8.1 and the claim hierarchy (Table 5, Section 8) convert a previously qualitative "paper computer" framing into a formally structured hypothesis with Level-0 through Level-3 claims clearly separated. A central contribution is that it provides a set of structural tests whose results could meaningfully compare the PPH against cipher-text and pseudo-text null hypotheses without requiring semantic assignments in advance.

The framework does not solve the Voynich Manuscript. It solves a narrower, prerequisite problem: how to represent the manuscript losslessly, compute structural invariants reproducibly, and test candidate generator classes without importing semantic assumptions. The next scientific question is not whether the harmonic core translates the text—it provably cannot—but whether the manuscript's structural profile can exclude natural-language, cipher, pseudo-text, or procedural-generator families under preregistration-ready controls.

Representation is solved. Interpretation remains conditional.

Code, benchmarks, and data should be deposited in a public repository before external review or formal submission to support reproducible Voynich scholarship.

Acknowledgments

The author thanks the Beinecke Rare Book & Manuscript Library at Yale University for maintaining public access to high-resolution scans of MS 408. The author also thanks the open Voynich scholarship community for maintaining public transcription files (EVA interlinear, Stolfi files) that make reproducible computational analysis possible.

Reproducibility Statement

All code implementing the pipeline, harmonic-core computation, cipher benchmarks, and EVA structural analysis should be made publicly available before external review or formal submission. Experiments use Python 3.10+, NumPy, and SciPy (exact version requirements in `requirements.txt`). Random seeds: all experiments use `numpy.random.seed(42)`

and `random.seed(42)`. Corpus sources: six Project Gutenberg public-domain prose passages (full text in `corpus.py`). EVA transcription source: uploaded IVTFF EVA archive `LSI_ivtff_0d(1).txt`, using only Takahashi `;H` lines and page metadata tags for illustration type, Currier language, and hand. Proxy frequency diagnostics, where retained for smoke testing, use top-40 token frequencies from Landini 2001 and Montemurro and Zanette 2013. Preprocessing: all English control text uppercased with non-alphabetic characters removed; EVA word tokens cleaned of editorial comments, uncertain filler symbols, and line/paragraph separators. Self-loops excluded from incidence matrix per Remark 4.2. Rare-token handling: no merging; all tokens present in EVA frequency table retained.

References

Barbarossa, S., Sardellitti, S., and Ceci, M. (2020). Topological signal processing over simplicial complexes. *IEEE Transactions on Signal Processing*, 68:2992–3007.

Beinecke Rare Book & Manuscript Library, Yale University (2024). Voynich Manuscript. `https://beinecke.library.yale.edu/collections/highlights/voynich-manuscript`. Accessed 2025.

Currier, P. H. (1976). New research on the Voynich Manuscript: Preliminary report. Seminar presented at the National Security Agency, November 1976. Declassified and widely circulated; reproduced in D'Imperio [1978].

Davis, L. (2020). Statistical and paleographic evidence for a bipartite scribal structure in the Voynich Manuscript. *Preprint*, available via `https://voynich.nu`.

Defferrard, M., Bresson, X., and Vandergheynst, P. (2016). Convolutional neural networks on graphs with fast localized spectral filtering. *Advances in Neural Information Processing Systems*, 29.

D'Imperio, M. E. (1978). *The Voynich Manuscript: An Elegant Enigma*. Aegean Park Press.

Greshko, M. A. (2025). The Naibbe cipher: A medieval cipher mechanism consistent with Voynich-like statistical structure. *Cryptologia*. DOI: 10.1080/01611194.2025.2566408.

Hodgins, G. W. L. (2011). Accelerator mass spectrometry dating of the Voynich Manuscript. University of Arizona AMS Facility. Report commissioned by the Beinecke Library.

Jiang, X., Lim, L.-H., Yao, Y., and Ye, Y. (2011). Statistical ranking and combinatorial Hodge theory. *Mathematical Programming*, 127(1):203–244.

Landini, G. (1998). Evidence of linguistic structure in the Voynich Manuscript using spectral analysis. *Cryptologia*, 26(4):275–295.

Landini, G. (2001). Evidence of linguistic structure in the Voynich Manuscript using spectral analysis. *Cryptologia*, 25(4):275–295.

Leimkuhler, B. and Reich, S. (2004). *Simulating Hamiltonian Dynamics*. Cambridge University Press.

Lim, L.-H. (2015). Hodge Laplacians on graphs. *SIAM Review*, 62(3):685–715.

Markov, A. A. (1913). An example of statistical investigation of the text "Eugene Onegin" concerning the connection of samples in chains. Proceedings of the Academy of Sciences of St. Petersburg.

McCrone Associates (2009). Materials analysis report: Voynich Manuscript. Report archived at the Beinecke Rare Book & Manuscript Library, Yale University.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems*, 26.

Montemurro, M. A. and Zanette, D. H. (2013). Keywords and co-occurrence patterns in the Voynich Manuscript: An information-theoretic analysis. *PLOS ONE*, 8(6):e66344.

Reddy, S. and Knight, K. (2011). What we know about the Voynich Manuscript. In *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pages 78–86.

Sauer, T., Yorke, J. A., and Casdagli, M. (1991). Embedology. *Journal of Statistical Physics*, 65(3-4):579–616.

Stolfi, J. (2000). Interlinear Voynich transliteration files (EVA). `https://www.ic.unicamp.br/~stolfi/voynich/98-12-28-interln/`. Accessed 2025.

Takens, F. (1981). Detecting strange attractors in turbulence. In *Dynamical Systems and Turbulence, Warwick 1980*, Lecture Notes in Mathematics, 898:366–381. Springer.

Zandbergen, R. (2023). The Voynich Manuscript transliteration: Standards, formats, and the Super Transliteration Alphabet. In *Proceedings of the Voynich Transcription Workshop*, CEUR-WS.

A Anticipated Reviewer Objections and Responses

This appendix addresses anticipated objections to clarify the scope and claims of the framework. Its inclusion is not defensive but methodological: stating what a framework *cannot* do is part of specifying what it *does* do.

Objection 1: "This paper does not decipher the Voynich Manuscript." Response: Correct. The paper explicitly separates representation from decipherment (Remark 3.13, Section 8.1). The contribution is a lossless structural-analysis layer for generator-class discrimination, not a proposed decipherment. Papers claiming to decipher the manuscript that have not passed Tiers 1–4 of the falsification protocol should be evaluated on that basis.

Objection 2: "Circuit rank is too coarse a measure." Response: Correct, and stated explicitly (Remark 4.10). Circuit rank is the entry-level invariant. Section 4.4 and Table 1 define seven additional features (successor entropy, burstiness, spectral peaks, motif density, etc.) that provide finer-grained discrimination. Circuit rank is a baseline with a clean algebraic proof (Proposition 4.11), not the final word.

Objection 3: "The Hodge decomposition does not imply semantic content." Response: Correct, stated as the core of the paper (Remark 3.13). The harmonic core measures transition circulation; it is a structural, not semantic, fingerprint. This is a feature, not a bug: structural invariants that carry no semantic presuppositions are more broadly applicable and harder to overfit than ones that do.

Objection 4: "A substitution cipher preserves graph structure, so the framework cannot recover keys." Response: Correct, and *proven* as Proposition 7.1. The paper does not claim to recover substitution keys via the harmonic core; the Caesar and Vigenère benchmarks are explicitly framed as sanity checks confirming the theorem (Section 7). Key recovery for monoalphabetic ciphers is handled by the trigram LM, which achieves 100% accuracy at T1.

Objection 5: "The EVA experiments use a proxy, not the actual transcription." Response: This objection applied to an earlier draft. The present draft includes a full run on the uploaded IVTFF/Takahashi EVA archive (Section 9). Remaining limitations are not proxy dependence, but robustness: alternative transcriptions, glyph-level tokenization, matched natural-language baselines, and section-aware controls must still be tested.

Objection 6: "The preregistration claim is unverified." Response: The paper uses "preregistration-ready" throughout to indicate that the protocol is specified in advance and designed for preregistration, but that a formal deposit (OSF/Zenodo/GitHub timestamped release) has not yet been made. This deposit should occur before external review, formal submission, or any stronger claim that the protocol is officially preregistered, as noted in Section 6.

Objection 7: "The Pasigraphic Procedural Hypothesis is speculation." Response: Correct. The PPH is explicitly labeled hypothetical in Table 5 and Definition 8.1. The paper's structural framework is neutral between the PPH and competing hypotheses; it provides tools to test the PPH, not evidence that it is true.