Shrinking the Constraint Envelope of the Voynich Manuscript: A Firewall-Disciplined, Self-Correcting Computational Program

Tim Walsh
Principal Researcher, DireLabs, LLC
tim@direlabs.com direlabs.com ti.mims.ms

Preprint v7 — 25 July 2026

Abstract

We report a computational program on Beinecke MS 408 (the Voynich Manuscript) whose goal is not decipherment but the defensible *shrinking of the constraint envelope* around what the book is: which classes of encoding, language, and community of origin remain consistent with the evidence, with every claim carrying an explicit A–D evidence grade. Methodologically, the program couples three disciplines: a synthetic ground-truth *harness* (real-language, cipher, and null-generator controls) that every method must pass before touching the real manuscript; a *firewall* in which all statistics are produced by deterministic, versioned code writing machine-readable results; and a standing *clean-context adversarial refutation pass* applied to every finding. We confirm the well-known low-entropy anomaly (conditional character entropy $h_2 \approx 2.08$ vs 3.07–3.91 for medieval natural-language controls), a genuine two-system (Currier A/B) structure, strong positional grammar, and a dense paradigmatic morphology (edit-distance-1 main component 0.80). We find no lexical label→feature naming system for the herbal even at increased statistical power; verbose ciphers are excluded on morphology-network grounds while an abjad/abbreviation class is revived; and *no single statistic—nor a vector of them—localizes meaning*: against honest structured-meaningless baselines, a meaningless generator out-scores meaningful text on every axis we can measure. Building the *mid-level* linguistic layers under a null-model-correction framework, we find the manuscript's structure lives below the word: it lacks the natural-language surface content>function collocational gap and shows only weak distributional word-class structure ($\approx$0.13–0.19$\times$ real language), and the two Currier systems do not differ in this mid-level grammar once content is controlled. Turning that mid-level signature into a cryptanalytic discriminator, we *robustly exclude word-order-preserving ciphers* of real prose (1:1 substitution, abjad, nomenclator): they inherit the source language's strong surface syntax—universally across Romance, Germanic, and consonantal-Semitic controls, and by 6.8σ (function/content) and 8.0σ (word-class) once the null is deconfounded from topic drift—which the manuscript lacks. We initially over-reached to "the *entire* cipher-of-real-prose class," but engaging a concurrent hand-constructed verbose-homophonic cipher [Greshko, 2025] forced a *retraction*: the second leg of that argument (retained word-order information) rests on a statistic confounded by respacing and homophony, and verbose/homophonic ciphers of word-boundary prose sit at the edge of the manuscript's joint signature—neither robustly excluded nor robustly reached (*inconclusive* on our measures). So the manuscript is not a word-order-preserving cipher, but a Naibbe-style homophonic cipher is not ruled out; it and a template-driven / positional generative account both remain live. Pursuing that favoured class through five generations of generator, we map *how far* it reaches: a first family appears to leave the manuscript's morphological connectivity, lexical reuse, and frequency

law mutually incompatible with its low entropy, but this gross incompatibility turns out to be largely an artifact of a too-rigid morphology in which connectivity is welded to word length and alphabet size. Once connectivity is made an independent control (a larger character space, then variable word length), the incompatibility collapses to a single shallow, *coupled* frontier among character entropy, connectivity, and mean word length: a positional generator over a skewed type-lexicon comes within $\sim$0.03–0.05 on four of six distributional "hard" axes, but—under strict multi-seed scoring—*no single configuration reproduces the full set*, and the last obstruction has *moved* (from connectivity to an entropy/word-length wall) rather than vanished. The honest conclusion is a *retraction* of the earlier hard constraint, not a promotion to "no constraint": the distributional signature is far less constraining than the first pass implied, yet it still excludes word-order-preserving ciphers and still resists a clean joint fit, so within the one generative family we tested it under-determines the sub-mechanism without being uninformative. A candidate positive signal (a cross-organ root$\leftrightarrow$leaf visual regularity) was pursued to its measurement ceiling and is *unresolved and underpowered*. The most robust result is the architecture itself: across ten iterations, the refutation pass repeatedly overturned the program's own conclusions—including one produced by its own analysis code, a generative "sufficiency" claim overturned once its parameters were shown to have been fitted to the target, an over-strong *negative* ("no generative family reproduces the signature") walked back once the alleged obstruction proved a morphology artifact, and, most recently, the retraction of the over-strong cipher-exclusion above once we ran a concurrently-published cipher through our own discriminators—before any reached print. We make no translation or real-taxon claim of any kind.

Keywords: Voynich Manuscript; computational philology; constraint-based analysis; adversarial validation; evidence grading; reproducible research.

1 Introduction

The Voynich Manuscript (Beinecke MS 408) is an early-fifteenth-century codex written in an undeciphered script. Its computational study has a long record of confident, mutually incompatible "solutions," most of which pattern-match plausible output onto an unconstrained problem [Reddy and Knight, 2011, Bowern and Lindemann, 2021]. We take the opposite stance. We do not attempt to read the manuscript; we ask which *classes* of hypothesis the data can exclude or leave standing, and we require every claim to carry an evidence grade (A: replicated/robust; B: candidate; C: directional/weak; D: speculative). Decipherment is explicitly not a goal, and no output of this program should be read as a translation.

The program's premise is that the field's recurring failure is methodological: plausible-sounding results are generated faster than they are refuted. We therefore invert the usual order—validation first, claims second—and we make the validation machinery, not any single finding, the primary contribution.

2 Data and Transliteration

We use the standard Zandbergen–Landini (EVA) transliteration as primary, with the v101 (GC) transliteration as a sensitivity layer [Zandbergen, 2024]. All text analyses stratify by the two Currier "languages" A/B [Currier, 1976] and by scribal hand [Fagin Davis, 2020]. Visual features of the herbal illustrations were annotated through a structured vision-model pipeline emitting coarse categorical morphology (e.g. root colouring, leaf arrangement) under a strict schema; annotation reliability is measured and carried through every downstream test. Natural-language comparanda are period-appropriate Latin, German, Italian, and (consonantal) Hebrew herbals and texts.

The extant manuscript is also incomplete, which bounds every lexical statistic below. Of the foliated range f1–f116, 14 folio numbers are absent ($f12$, $f59$–$f64$, $f74$, $f91$–$f92$, $f97$–$f98$, $f109$–$f110$); 102 folios (207 inscribed sides) survive, so ≈88% of the foliated range is present and at least ≈12% is lost (a lower bound—folios lost before or after the surviving foliation cannot be detected). Because the vocabulary is far from saturating—a Heaps exponent $\beta = 0.73$, with ≈20% of the words on a page still novel even in the final quartile of the text (34,111 tokens, 7,424 types, type–token ratio 0.218)—the missing folios very likely held substantial unseen vocabulary, and the combination of high lexical richness and incompleteness compounds the sparsity that limits every anchor and label test below.

3 Method: Harness, Firewall, and Adversarial Refutation

Three disciplines are load-bearing.

Harness (validate before you claim). No method is trusted on the real manuscript until it behaves correctly on synthetic ground truth: real-language controls (H4), a verbose homophonic cipher (H2, Naibbe-class), a self-citation / auto-copying generator (H3) [Timm and Schinner, 2020], and parameterised null generators. A method that cannot separate these knowns is not used on the unknown.

Firewall (numbers come only from code). Every statistic is produced by deterministic, versioned code with fixed seeds, written to machine-readable JSON; each result records its producing script, git commit, dataset version, and parameters. No number in this paper is estimated or recalled—each is copied from a result file.

Adversarial refutation (a standing pass). Every finding is submitted to a clean-context skeptical reviewer instructed to build the strongest case against it, with positive/associational claims held to the harshest scrutiny (plausible-but-wrong positives being the field's main failure mode). A claim is graded A/B only after it survives this pass. As documented in §5, the pass has repeatedly overturned the program's own conclusions.

Null-model correction (mid-level probes). Distributional statistics of word structure are confounded by nuisance parameters—type-token ratio, sample size, the number of induced units—so raw values are not comparable across corpora. We therefore express every mid-level statistic (§4.5) as a z-score against an ensemble of null replicates that share the corpus's own nuisance parameters but destroy the linguistic signal (e.g. an order-shuffle that preserves the exact word multiset). A statistic is interpreted only after this correction calibrates on known corpora; several first-pass probes did not, and yielded no manuscript claim.

4 Results

4.1 Established structure [Grade A/B]

The conditional character-entropy anomaly replicates [A]. Voynichese second-order character entropy is $h_2 = 2.08$ (all words; 2.16 with spaces under the Lindemann–Bowern convention [Lindemann and Bowern, 2022]), far below matched medieval controls: German 3.07–3.22, Latin 3.15–3.23, consonantal Hebrew 3.91, while first-order entropy $h_1 \approx 3.87$ is unremarkable. The anomaly is robust across Currier A ($h_2 = 2.15$) and B (1.94) and across scribal hands.

Two statistically distinct systems (Currier A/B) [A]. Unsupervised clustering of pages by word co-occurrence recovers the Currier A/B division as the dominant structure (adjusted Rand index of the two largest clusters vs. Currier = 0.44 in EVA, 0.90 in v101)—the strongest structure recoverable from the text alone [Montemurro and Zanette, 2013].

A dense paradigmatic morphology [B]. The edit-distance-1 word network has a giant component covering 0.80 of types (Currier A 0.75, B 0.82), versus 0.16 for Latin and 0.22 for Italian—an order of morphological regularity unlike natural-language controls at matched size. Strong positional grammar (paragraph- and line-level regularities) likewise replicates.

Section↔text co-variation, Language A only [B]. Vocabulary tracks illustration section within Language A (within-dialect alignment ARI = 0.35) but not within Language B (0.004): the two systems differ not only in vocabulary but in how text relates to content.

4.2 What is disfavoured [Grade C]

No lexical naming system for the herbal [C]. An anchor hunt for word→feature associations returns no survivor, and a higher-power graded (count-based) re-run over the most powerful feature bands finds no association above chance (permutation $p = 0.48$ against a permuted-feature null; the raw multiple-testing "hits" are false positives). Label-adjacency anchoring is infeasible on the herbal (labels are page-unique). A power analysis bounds this: page-level anchors are recovered at effect size $\phi \geq 0.4$ but not below $\phi \leq 0.3$, so weak or rare-feature mappings are not excluded.

Verbose ciphers excluded; abjad/abbreviation revived [C]. A verbose cipher expands each plaintext letter into multiple glyphs, which *deductively* destroys the edit-distance-1 network (a Latin cipher's morphology main-component falls from 0.28 at 1:1 substitution to 0.001 under verbose expansion); homophone-rich verbose ciphers additionally collapse word-order information ($\Delta I = 0.013$ vs 0.356 for a type-preserving map). Conversely, an *abjad* of real Latin (vowel-dropping to consonantal skeletons) reaches the Voynichese morphology band (main component 0.90 vs the manuscript's 0.80)—so the dense network does not require a constructed morphology, and the abjad/abbreviation class [Hauer and Kondrak, 2016] is revived as a live structural lead *on the morphology axis alone*. §4.6 tests that revived class against the joint signature and excludes it on its function/content syntax (it is a word-order-preserving cipher).

4.3 What is not established / underdetermined

Meaning is not localizable by structural statistics. We attempted to place the manuscript on interpretable meaningful↔meaningless axes. Against *honest structured-meaningless* baselines, the meaningless generator out-scores meaningful text on every axis: word-order information for a block drift-null is $\Delta I = 0.57$ vs 0.36 for natural Latin, and the morphology network for a meaningless abjad is 0.89 vs 0.87 for a constructed language—with the manuscript sitting *below both* endpoints ($\Delta I = 0.31$, morphology 0.80). Every VLM-computable statistic we tried is a structure detector, not a meaning detector. *The manuscript is maximally structured on every axis we can measure, yet no axis separates meaning.*

No encoding family is distinguished [C]. Under equal, held-out tuning with de-collinearised metrics, no generative family (cipher, self-citation, constructed language, abbreviation, abjad, and compositions thereof) is robustly closest to the manuscript; a whitened (Mahalanobis) re-analysis agrees but is itself regularisation-dependent (covariance condition number $\approx 1.0 \times 10^4$), which only reinforces that the encoding "bracket" is a descriptive ordering, not evidence for any family. §4.6 sharpens this only for the *word-order-preserving* ciphers of real prose, which it excludes; verbose/homophonic ciphers are not excluded.

4.4 A candidate positive, resolved to unresolved: the root↔leaf bundle

The one associational signal that survived early scrutiny was a cross-organ visual regularity: a plant's root colouring co-varying with its leaf arrangement (Cramér's $V = 0.40$). Its history is

a cautionary tale in miniature. A first re-annotation model failed to reproduce it ("single-model artifact"); a third model reproduced it and it survived stratification by scribal hand and dialect and a palette-style control (within-stratum permutation $p = 0.003/0.014$); we then added a genuinely out-of-lineage rater (a non-Anthropic frontier model whose root labels agree 0.86–0.91 with the others), and it did *not* reproduce the association—no cross-rater pairing survives. Crucially, the reproduce/not-reproduce split does not track vendor lineage, and the leaf feature is noisy for every rater (pairwise agreement ≈ 0.45). On the high-confidence consensus subset (raters agreeing on both organs) only $n = 74$ pages remain, with minimum detectable effect $\phi \approx 0.33$—underpowered for the ~ 0.28 effect, and where labels are reliable the association is $\approx$ nil ($V = 0.19$, $p = 0.63$). **The honest terminal status is unresolved: the feature is too noisy for model annotation to adjudicate; a pre-registered human panel is the only decisive test.** We record the whole arc, including its flip-flops, rather than report whichever model was added last.

4.5 Mid-level structure: where the structure lives [Grade C]

Having shown that magnitude-of-structure statistics cannot reach meaning, we built the *mid-level* layers between structure and semantics—collocation, word-classes, morphology—each null-corrected (§3) and stratified by Currier A/B.

No natural-language surface function/content organisation [C]. Real languages show a content>function collocational gap: content words have specific collocates (peaked neighbour distributions) while frequent function words are flat. Null-corrected against an order-shuffle baseline, this gap is strongly present in real controls (Latin $z = 19.1$, German $z = 8.5$) but *absent* in the manuscript ($z = -1.3$ for Currier A, -4.8 for B): its content-band words carry only near-chance collocational selectivity, and its most-frequent words are, if anything, the *more* collocational—template-like local repeats rather than flat function words. The result is robust across function/content band-cutoff choices.

Weak distributional word-class structure [C]. Inducing word classes by clustering context vectors and measuring adjacent-class transition information [Brown et al., 1992], again null-corrected, real languages are strongly class-structured ($z \approx 14.8$) while the manuscript is only $\approx$0.13–0.19$\times$ that level in both dialects—weak part-of-speech organisation, robust across the number of classes and context-vocabulary size.

The two dialects do not differ in mid-level grammar (once content is controlled) [C]. A global A-vs-B contrast suggested Currier B carries more word-class structure than A; but Currier A/B is confounded with section (A dominates the herbal, B the biological/recipe quires). The herbal is the only section containing both dialects, giving a content-controlled comparison: there the apparent difference *reverses* (word-class z: global $B - A = +1.79$; within-herbal $B - A = -1.24$). So the A/B mid-level difference is a section/content effect, not a dialect property—the two systems' established difference is lexical, not grammatical (this within-herbal contrast is itself underpowered, Herbal-B $\approx$ 3.3k tokens). *Scope (L7): these are surface-distributional results; a verbose/abbreviatory cipher or a template-driven generator would depress surface word-syntax while leaving character and morphology structure intact, so they do not prove the absence of an underlying grammar.*

4.6 Ciphers of real prose: word-order-preserving excluded, homophonic *not* [Grade B / C]

The mid-level syntax measures (§4.5) double as a cryptanalytic discriminator, because they behave differently under enciphering than under generation. We assembled a *joint signature*—character entropy h_2, word-order information ΔI, edit-distance-1 morphology, Zipf slope, and the two null-

corrected mid-level z-scores—for the manuscript and for candidate ciphers of real Latin, and asked which reproduces it. *This subsection records both a robust result and a retraction: an earlier version of this program claimed the* ***entire*** *cipher-of-real-prose class excluded; engaging a concurrent, hand-constructed cipher forced us to withdraw that to a narrower claim.*

Word-order-preserving ciphers are robustly excluded [B]. A 1:1 substitution, a vowel-dropping abjad, and a light nomenclator applied to Latin all inherit its strong surface syntax: function/content $z = 13.8/21.0/9.8$ and word-class $z = 11.3/13.0$, versus the manuscript's weak $z \approx -1.2/1.9$ (Currier A), $-4.7/2.6$ (B). The *gap* is large and stable: it is *language-universal* (no natural control is weak—Latin $z = 20.0/14.9$, Italian 8.2/11.4, German 8.6/14.1, and, decisively for the abjad hypothesis, consonantal Hebrew 9.7/4.1), and once the measures' true seed-noise is used (block-bootstrap-with-replacement being invalid here—duplicate blocks inflate collocation), an order-preserving cipher sits 6.8σ (function/content) and 8.0σ (word-class) from the manuscript, well outside its own uncertainty. We stress that the exclusion rests on this *gap*, not on the manuscript's absolute value: whether the manuscript's weak word-class z is itself grammar or drift is unsettled—a deconfounding null (within-section shuffle) leaves it essentially unchanged ($1.98 \to 1.97$, arguing grammar), but an earlier control (real Latin word-*types* under a block wrapper) reproduced the same weak-positive value with no grammar at all (arguing drift), and its own subsample CI, $[-1.44, 1.94]$, crosses zero. The order-preserving ciphers sit far above that entire interval either way. So word-order-preserving ciphers of real prose—the abjad/abbreviation class §4.2 had revived on morphology alone—are firmly excluded, independently of the two-point band.

But the broader "entire cipher-of-real-prose class" claim is retracted [C]. The earlier version of this program excluded *homophonic* and transposition ciphers too, via a second leg: they produce weak syntax but were argued to destroy the manuscript's retained ΔI (which we read as block/section structure, §4.1). A concurrent publication forced a re-examination. Greshko's *Naibbe* cipher [Greshko, 2025] is a hand-constructable, 15th-century-plausible, *decipherable* verbose-homophonic substitution that encrypts Latin into Voynich-like ciphertext and reproduces many of its statistics. Running its actual ciphertext through our discriminators first looked like confirmation—its ΔI collapses to 0.004—but that reading does not survive scrutiny. The collapse is an artifact of two things unrelated to "being a cipher." First, *respacing*: Naibbe refragments the Latin into non-word units before encryption, and real *word-boundary* Latin (one book of the same source) already has $\Delta I = 0.076$, *inside* the manuscript's band—on this text and spacing scheme, most of Naibbe's ΔI loss ($\approx$80%) is attributable to the plaintext respacing, before the cipher runs (a same-text decomposition on the full compilation is not available, as only fragments survive). Second, *homophony*: with word order and boundaries fully preserved, ΔI falls monotonically as homophones-per-type rises (0.076 at 1 to 0.005 at 32). So ΔI is a homophony/type-token-coupling detector, not the clean "retained word-order" axis the second leg assumed; it is retired *as applied*. We then closed the one remaining form of that leg—a *block-scale, like-for-like* ΔI (matched token budget and partition scales, no respacing, null-corrected), cast as the mutual (in)compatibility of low character entropy and block-scale word-order information on the (h_2, ΔI_{block}) plane. Sweeping verbose×homophony under a fair, in-alphabet homophony model, *no* configuration reaches the manuscript's corner (closest normalised distance 1.36, and only at $\sim 3\times$ its word length): matching its block-scale ΔI forces h_2 above its window, and vice versa. So the block-scale plane *weakly separates* verbose+homophonic ciphers—it does *not* collapse—but this does not revive ΔI into a hard, standalone discriminator (one soft axis, a single generator family), so the class verdict below is unchanged. (A first pass of this test reported the opposite—a cipher *reaching* the corner—until the refutation pass traced that to a homophony model whose marker glyphs artificially deflated h_2; see §5.) When we instead sweep verbose+homophonic ciphers of *blocked, word-boundary* Latin (multi-seed), the result is *inconclusive*: the class is neither robustly excluded (its word-class z sits 1.2σ from the manuscript,

a non-separation by a hair) nor robustly reached (across 8 parameterisations × 8 seeds, exactly *one* seed-configuration touched all four bands, and the discriminating z-scores are unstable, spread $\sim$4). An independent joint-signature analysis [Parisel, 2026] likewise finds that simple positional/grille generators fail to reproduce the joint signature, and reads the residual as cipher-like—adjacent to our position, though we keep a generative account equally live. **We therefore retract "the entire cipher-of-real-prose class is excluded" to "word-order-preserving ciphers are excluded; the verbose/homophonic class is inconclusive (neither excluded nor robustly reached)"**—so a Naibbe-style cipher is *not* ruled out by our measures, which is why we treat it as a live hypothesis rather than a refuted one. *No decipherment or meaning claim is made or implied (L7); this is a statistical statement about which cipher classes reproduce the distributional signature.*

4.7 The favoured generative class: an apparent constraint that dissolves [Grade C]

Having excluded *word-order-preserving* ciphers (§4.6) and, at the time, taken the template/positional generative reading as the residual, we tested that class directly—can such a generator, *not* tuned to the manuscript, reproduce the *full* eight-axis signature? The answer, after a refutation pass overturned a first-pass positive, is an informative *negative* that sharpens the constraint.

A context-free positional generator is insufficient. We built a three-ingredient generator—a positional character grammar (disjoint small glyph inventories per slot), a block-theme wrapper, and context-free slot-filling—whose parameters were swept over an a-priori grid rather than fitted. It reproduces the manuscript's character entropy ($h_2 = 2.14$), block-scale ΔI (0.08), and (via block contrast) its weak-but-positive word-class structure, but across a 64-point a-priori grid it does not *simultaneously* match the manuscript's morphology-network connectivity (edit-distance-1 main component ≈ 0.75; the generator runs high, ≈ 0.97–1.0), its lexical reuse (type-token ratio ≈ 0.22; generator ≈ 0.59), or its frequency slope. These are not proven unreachable—connectivity and slope move toward band at the grid's edge—but they are *coupled*: the branching that lowers connectivity raises entropy out of band, and the concentration that lowers the type-token ratio does likewise, so the context-free "bag of slots" cannot place them in-band together. Two of the eight axes carry little independent weight here: the function/content and word-class z-scores are scored against two-point Currier A/B ranges (not bootstrap intervals), and a control shows the manuscript's weak-positive word-class z is reproduced by real Latin word *types* sampled context-free under the same block wrapper ($z_{fc} = -1.4$, $z_{wc} = 2.5$)—so that "positive" signal is sectional vocabulary drift, not a grammatical feature a generator must earn, and we do not lean on it. *Process note (a self-correction on the record):* a first-pass verdict had graded this generator a "class sufficiency" at grade B; the refutation pass showed its constants had been grid-selected to hit the manuscript's bands (a fitted point, not an a-priori draw) and that its weak-syntax test used a one-sided threshold a full order shuffle also passes—and the corrected, band-honest analysis turned the apparent positive into this negative.

Adding word reuse helps but hits a Pareto tension. The three missed axes are all frequency-concentration failures, so we added a word-reuse (copy-from-history) mechanism and swept it a-priori (104 configurations, including global preferential attachment). Reuse individually rescues every missed axis—connectivity in-band from copy-rate $\rho \geq 0.6$, reuse from $\rho \geq 0.4$, frequency slope from $\rho \geq 0.8$—and global preferential attachment additionally *restores* the manuscript's weak-positive word-class structure (z_{wc} from -1.0 up into band, though as just noted that axis is soft). But no configuration satisfies even the whole frequency group at once, let alone jointly with the entropy/ΔI group (the grid's ceiling is four of eight axes): the copy rate that concentrates frequency simultaneously deflates the character entropy and the block-scale ΔI. So token-level copying does

not, at the swept frequency exponent and block length, concentrate frequency without spending entropy and ΔI.

Type-level concentration resolves that trade-off but leaves a residual coupling. The natural fix is to concentrate frequency at the *type* level—a genuinely small, skewed lexicon of positional-grammar words, sampled context-free—rather than by copying tokens. Swept a-priori (144 configurations over lexicon size, branching, frequency exponent, and block contrast), this *does* achieve what token reuse could not: the type-token ratio now sits in band *jointly* with the character entropy, dissolving the entropy-versus-reuse tension. Yet the full signature is still not reached (ceiling again four of eight): the obstruction migrates to *morphological connectivity*— matching the manuscript's edit-distance-1 connectivity (≈ 0.75) forces a lexicon/branching regime that is out of band on entropy, reuse, and frequency slope— together with a block-contrast trade-off (retained ΔI favours weak block contrast while a positive word-class z favours strong contrast). Taken alone, this looked like a constraint—across those three families none reproduced the manuscript's full signature.

But that apparent coupling was largely an artifact—the walk-back. A clean-context refutation flagged that in every generator so far, edit-distance-1 connectivity is rigidly tied to word length and branching: a small (low-entropy) character alphabet gives a dense edit-space, so connectivity saturates near 1.0 and can only be lowered by enlarging the alphabet, which raises entropy. That is a property of the *morphology parameterisation*, not of the manuscript. Making connectivity an *independent* control dissolves most of the coupling—but does not eliminate it, and we are careful not to over-read the reversal. First (multi-seed scoring: an axis "matches" if its min–max range over the seeds overlaps the manuscript's band—a deliberately lenient rule, see Limitations), enlarging the character space with a connected-core-plus-isolates lexicon lets the edit-distance-1 component vary freely: it overlaps the manuscript's band together with ΔI, lexical reuse, and frequency slope, leaving a shallow *entropy-versus-connectivity* near-miss (with entropy in band, connectivity reached only ≈ 0.63 against the manuscript's ≈ 0.75)—though *no single configuration* matched all five non-trivial axes (0 of 48; ceiling four of five). Second, adding *word-length variance*—words of different lengths connect only by an insertion/deletion, so spreading lengths thins the edit-graph without enlarging the alphabet—brings connectivity squarely into band (≈ 0.76). But again *no single configuration* reproduces the full hard-axis set (0 of 48; ceiling four of six; none even at five of six): at the configuration where connectivity enters the band, the residual is *two* misses, not one—a ~ 0.03 overshoot on entropy *and* a mean word length of ≈ 6.4 against the band's ≈ 5.1. We conjecture the length overshoot is a construction artifact (a small alphabet saturates its few short words), but that conjecture is *untested*, and it may not be separable: the in-band-connectivity configurations are structurally confined to the smallest alphabet with the widest length spread, so the length overshoot may be the same entropy/connectivity wall reappearing in a third coordinate. The honest reading is therefore a *retraction of the first-pass "hard constraint," not a promotion to "no constraint."* The i08 gross multi-axis incompatibility was largely a morphology-parameterisation artifact; freeing connectivity collapses it to a single shallow, coupled entropy–connectivity–length frontier that a flexible generator approaches to within ~ 0.03–0.05 on four of six axes—yet even this five-to-six-knob family does not jointly reproduce the signature, so the fit is *incomplete, not vacuous*. Within this one generative family the hard axes therefore under-determine the sub-mechanism (lexicon size vs. isolate fraction vs. length spread); they do *not* license the stronger claim that the distributional signature is uninformative about mechanism *in general*—indeed the same signature excludes the word-order-preserving ciphers (§4.6). The load-bearing constraints stay that (now narrowed) exclusion and the qualitative character/morphology structure (§4.1). (Grade C throughout; no sufficiency of a specific mechanism and no identification are claimed; two axes—the function/content and word-class z-scores—are soft two-point ranges we do not count; L7.)

5 Discussion

Two substantive conclusions and one methodological one. Substantively: (i) the manuscript is a genuine, richly structured, two-system formal script whose morphology network is reachable only by a length-preserving script over an already-structured inventory—which revived the abjad/abbreviation class, until the joint-signature test (§4.6) excluded it, together with the other word-order-*preserving* ciphers, on their strong surface syntax (the *verbose/homophonic* class is left inconclusive—§4.6); (ii) whether that structure carries *meaning* is underdetermined, and we show *why* it is hard: the statistics that quantify structure (word-order information, character entropy, morphological density) do not discriminate meaning, because structured-meaningless processes reproduce or exceed them. This reframes the central question from "what does it say" to "how far along each interpretable dimension does it sit," and marks the honest boundary at which purely distributional methods stop. (iii) The mid-level probes (§4.5) locate *where* the structure lives: strongly at the character (entropy anomaly) and morphology (edit-distance-1 network) levels, but *weakly at the word level*—no natural-language function/content collocation, weak word-class organisation. The cryptanalytic reading of that profile (§4.6) yields a robust negative—a system so syntactically thin above the word cannot be a word-order-*preserving* cipher of real prose (those keep the plaintext's strong syntax, a 6.8–8.0σ deconfounded gap)—but *not* the sweeping "no cipher of real prose" we first claimed: a verbose-homophonic cipher weakens surface syntax too, and (once one does not fault it for the respacing/homophony artifact that flattens ΔI) sits at the edge of the manuscript's joint signature. Both a homophonic cipher and a template-driven/positional generative process therefore remain live. Pursuing that favoured account across five generations of generator (§4.7) then taught a lesson about its own limits: what first looked like a joint-signature constraint (successive mechanisms each dissolving one blocker and exposing another) turned out to be largely an artifact of a rigid morphology in which connectivity is welded to word length and entropy—and once connectivity is given an independent control (a larger character space, then variable word length), the gross multi-axis incompatibility collapses to a single shallow coupled frontier that a flexible generator approaches to within $\sim$0.03–0.05 on four of six hard axes, though no single configuration reproduces the full set. The correct inference is bounded: *within this one generative family* the hard axes do not pin down the sub-mechanism, so i08's "hard constraint" is *retracted*—but this is not the stronger claim that the distributional signature is uninformative in general, which would contradict the word-order-preserving-cipher exclusion two subsections above (§4.6), where the *same* signature decisively separates classes. The tightening of the envelope thus comes from that cross-class exclusion and the qualitative below-the-word structure, not from a joint-fit barrier within the favoured class. That the two Currier systems share this mid-level profile (their difference being lexical, not grammatical) further constrains "two languages" readings toward "two registers/vocabularies of one system."

Methodologically, the architecture is the result. The clean-context refutation pass overturned, before publication, several of the program's own most attractive conclusions: a "meaning certificate" read of the word-order statistic; a "constructed-language best fit"; a "needs constructed morphology" claim (refuted by the abjad control the reviewer demanded); a "whitening confirms" claim (withdrawn once the covariance was shown ill-conditioned); a set of false-positive anchors caught only by a mandatory null control; a positive root$\leftrightarrow$leaf verdict, first over-claimed and then over-killed, corrected to unresolved; an apparent Currier A-vs-B grammatical difference that a dedicated content-controlled study overturned as a section confound; a first-pass "the manuscript is favoured as a generation process" read of the joint-signature test, rejected as circular (its only match was a Voynich-tuned generator) and narrowed to the robust negative—after which the refutation's own demanded control (typologically diverse languages) *upgraded* that negative from Latin-specific to universal; and, most recently, a "the favoured generative class is sufficient" claim (grade B) overturned when the refutation

showed the generator's constants had been *grid-selected to hit the manuscript's own bands*—a fitted point mislabelled a-priori, the very circularity flagged one iteration earlier—and its weak-syntax test used a threshold a random shuffle also passes, so the corrected, band-honest analysis reduced it to a precise negative; and, most tellingly in the reverse direction, that same precise negative—"no tested generative family reproduces the joint signature"—was itself *walked back* a step later, when a refutation-directed generator (connectivity decoupled from entropy and word length) showed the alleged obstruction was largely a morphology-parameterisation artifact, shrinking a gross multi-axis incompatibility to a shallow coupled frontier the generator approaches to within ~ 0.03–0.05 on most axes; a subsequent refutation of *that* walk-back then curbed its over-reach in turn (no single configuration reproduces the full set), leaving the honest "retraction, not reversal." The pipeline over-read a negative, corrected it, and corrected the correction. Most consequentially, when a verbose-homophonic Voynich-imitating cipher was published concurrently [Greshko, 2025], our *own* first-pass read of its ciphertext scored as *confirmation* of our cipher exclusion—until the refutation pass showed that read was an artifact of the cipher's respacing and homophony (both flatten the word-order statistic with word order intact), forcing us to retract "the entire cipher-of-real-prose class is excluded" to the narrower, robust order-preserving claim (§4.6). The clean-context adversary overturned the program's own favourable first-pass error before it was acted on. The pattern recurred once more, in the reverse direction, on the follow-up test (§4.6) that closed the block-scale ΔI question: a first pass found a verbose-homophonic cipher *reaching* the manuscript's (h_2, ΔI_{block}) corner and concluded the word-order leg was dead even at block scale—until the refutation pass showed the "corner" was an artifact of a homophony model whose marker glyphs deflated h_2 by ≈ 0.34 bits; under a fair in-alphabet model the plane instead *separates*, reversing the finding's direction. Several mid-level probes were themselves refused for want of a clean null rather than reported as weak evidence. A pipeline that declines to over-read a negative as readily as a positive, and that can overturn a verdict asserted by its own code or its own prior iteration, is, we argue, the transferable contribution to undeciphered-corpus research.

6 Limitations

Visual features are model-annotated and noisy (leaf-arrangement inter-rater agreement ≈ 0.45), which caps power on any imagery-based test; several conclusions rest on a single manuscript-length corpus and are correspondingly power-limited; the extant text is itself incomplete ($\approx 12\%$ of the foliated range is lost and the vocabulary is non-saturating, so the missing folios likely held unseen types), which bounds every lexical and anchor statistic below; all vision raters to date share broadly similar training lineages, so "independence" is cross-vendor, not absolute; the cipher exclusion (§4.6) is now *partial*—robust for word-order-preserving ciphers (a large, deconfounded syntax gap), but the verbose/homophonic class is *inconclusive* (neither excluded nor robustly reached; one seed-configuration of 64 touched all four bands), so we make no whole-cipher-class claim; and the discriminating mid-level z-scores, while stable to seed noise, have a manuscript-side subsample CI that crosses zero and are measured on a single manuscript with no generator-side error model beyond seeds; the generative-fit result (§4.7) is a partial *deflation* whose own limits we flag plainly—(a) *no single configuration* matches all hard axes in either experiment (E25: 0 of 48, ceiling four of five; E26: 0 of 48, ceiling four of six, none at five of six), so "comes close on four of six" is the claim, not "reproduces all"; (b) the E25/E26 matching rule (multi-seed min–max overlap) is more lenient than the single-seed hard threshold used for the i08 negative it walks back, and was *not* applied back to the i08 families, so that comparison is not strictly apples-to-apples (≈9–15% of the "matched" axis-cells have their median just out of band); (c) the mean-word-length overshoot is *conjectured* to be a construction artifact but is untested and may not be separable from the entropy/connectivity

wall; (d) the connectivity target (≈ 0.75) is a 10k-token subsample value, against 0.80 for the full manuscript (§4.1); and (e) resampling is multi-seed but finite ($K \leq 6$) over just two hand-built generator families. Two of the eight axes (the function/content and word-class z-scores) are soft two-point dialect ranges we do not count. And text results are primarily EVA-based, with v101 as a sensitivity layer rather than a full replication. Grades reflect these limits.

7 Future Work

The most consequential open question the cipher retraction (§4.6) leaves is whether a verbose-homophonic cipher of real prose can be brought *robustly* into the manuscript's joint signature rather than only to its edge; the direct test is to engage the published Naibbe construction [Greshko, 2025] on its own terms—encrypting *word-boundary* (not respaced) prose, and measuring the specific character- and word-structure properties that construction claims—and to compare against the concurrent positional/grille analysis [Parisel, 2026]. The decisive next test for the root↔leaf question is a pre-registered independent *human* panel on the high-confidence consensus subset, with a power analysis reported first; more models cannot settle a noise-limited question. The generative-fit result (§4.7) has largely run its course: the i08 hard constraint is retracted and the residual is a single shallow coupled frontier, so the productive direction is no longer more generators but the clean-ups that would settle whether the full signature is *jointly* reproducible—eliminating the mean-word-length construction artifact and replacing the single-fixed-length assumption, re-scoring the i08 families under the same multi-seed rule for an apples-to-apples comparison, and putting the soft function/content and word-class axes on a proper generator-side null rather than a two-point range—after which the bounded headline ("within the favoured generative family, the distributional hard axes under-determine the sub-mechanism") can be stated with confidence, without over-reaching to a general non-discrimination claim the word-order-preserving-cipher exclusion already refutes. The mid-level results additionally invite a larger content-matched A-vs-B sample (the within-herbal contrast is underpowered) to firm up the "no grammatical dialect difference" finding, and a long-range/hierarchical dependency probe to complete the mid-level layer. None of these targets translation; each targets the grammar of the system, which is what the evidence can bear.

Reproducibility and open-source release

All statistics are produced by deterministic, versioned code (fixed seeds) writing machine-readable JSON to `results/`; every reported number records its producing script, git commit, dataset version, and parameters. The pipeline is released as an open-source package (`ms408`, Apache-2.0) with three things a reader can run directly: a public `evaluate(tokens)` entry point that scores any word-token stream—a candidate cipher output, a generator, a transliteration variant—against the manuscript's discriminator bands and returns a per-axis verdict *with each axis's caveat attached* (the confounded ΔI, the token-sensitive type–token ratio, and the soft mid-level syntax axes are flagged and not counted, so the tool cannot be quoted without its hedges); a `verify` command that reproduces the shipped reference bands from code; and the archived clean-context refutation briefs that document the self-corrections above. External corpora are never redistributed—each is fetched at run time from a pinned, checksummed, licence-annotated registry (consume-only). Matching the manuscript on these axes is *necessary, not sufficient*: the tool is built to withhold, not to confirm.

References

Claire L. Bowern and Luke Lindemann. The linguistics of the voynich manuscript. *Annual Review of Linguistics*, 7:285–308, 2021. doi: 10.1146/annurev-linguistics-011619-030613.

Peter F. Brown, Peter V. deSouza, Robert L. Mercer, Vincent J. Della Pietra, and Jenifer C. Lai. Class-based n-gram models of natural language. *Computational Linguistics*, 18(4):467–479, 1992.

Prescott H. Currier. Papers on the voynich manuscript. Technical report, Unpublished; collected in the New Signals seminars, 1976. Origin of the Currier A/B language distinction.

Lisa Fagin Davis. How many glyphs and how many scribes? digital paleography and the voynich manuscript, 2020. Manuscript Studies; identification of multiple scribal hands. UNVERIFIED details.

Michael Greshko. The Naibbe cipher: a substitution cipher that encrypts Latin and Italian as Voynich manuscript-like ciphertext. *Cryptologia*, 2025. doi:10.1080/01611194.2025.2566408; ciphertext at github.com/greshko/naibbe-cipher.

Bradley Hauer and Grzegorz Kondrak. Decoding anagrammed texts written in an unknown language and script. *Transactions of the Association for Computational Linguistics*, 4:75–86, 2016.

Luke Lindemann and Claire L. Bowern. Character entropy and the voynich manuscript. *Working paper*, 2022. Conditional character entropy (h1/h2) with spaces. details.

Marcelo A. Montemurro and Damian H. Zanette. Keywords and co-occurrence patterns in the voynich manuscript: An information-theoretic analysis. *PLoS ONE*, 8(6):e66344, 2013. doi: 10.1371/journal.pone.0066344.

Christophe Parisel. Evidence of layered positional and directional constraints in the Voynich manuscript: Implications for cipher-like structure, 2026. arXiv:2604.19762.

Sravana Reddy and Kevin Knight. What we know about the voynich manuscript. *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH)*, pages 78–86, 2011.

Torsten Timm and Andreas Schinner. The voynich manuscript: Symbol roots, features, and a self-citation generative method, 2020. Self-citation / auto-copying generation hypothesis. details.

René Zandbergen. The voynich manuscript (voynich.nu), 2024. EVA and ZL transliterations; IVTFF format. `https://www.voynich.nu`.