The Directed, Context-Conditioned Layer of Phonological Correspondences

Letting sound laws emerge from documented forms — a tensor-and-MDL account (protoform-agnostic)

Alejandro Toledo Martínez — Independent researcher (ORCID: 0009-0000-1277-9697)

Reframe note (current). The paper now stands on the programme's cryptographic instrument (`docs/cryptographic-framing.en.md`; §0.7): the per-pair correspondence is a *substitution key*, and Paper 2 asks whether that cipher is mono- or polyalphabetic. Two experiments have landed on data in hand — 2A (make-or-break context test, §0.5/§1) and 2B (the *unicity distance* of the key, §1.1). Results persist to a dedicated Paper-2 database, `data/db/paper2.db`. Split note. The reconstruction-free *point-cloud / pairwise-distance* result (systems placed by their correspondence behaviour; genealogy and areality emergent) was split off into its own paper and repo (`languages-point-cloud`). This paper keeps the *directed, context-conditioned* layer: NMF bundles of X_{ijo} and MDL emergent laws over X_{ijoc}.

0. Guardrails — pre-registered before any computation (the lessons from Paper 1's four review rounds)

Paper 1 was published-defensible only *after* four adversarial rounds exposed, in sequence: an invalid significance claim (a σ from 3 permutations), a fatal imputation artefact (52.6 % of a matrix was a fill constant that manufactured the headline "signature"), a novelty that vanished against trivial baselines, an over-read areal story, and a hypothesis (diversification "fingerprint") that outran its data. We encode each of those as a binding design commitment here, so Paper 2 is built defensively rather than repaired. No result below is reported until it clears these.

- G1 · Never impute; handle missingness explicitly. X_{ijo} / X_{ijoc} are sparse. Distinguish a *structural zero* (operator impossible in that context) from an *unobserved cell* (no data for that pair). Use masked / weighted factorization (missing cells excluded from the loss), never zero-filling or mean-filling. Every reported factorization states its coverage and the mask. *(This is the 52.6 % lesson, in tensor form.)*
- G2 · Baselines before novelty. No bundle or "law" counts until it beats trivial baselines on the *same* data: the pilot's symmetric order-0 repertoire; raw operator frequency; the per-language marginal; plain co-occurrence. Paper 1 showed branch recovery is *generic* (edit distance ties it) — Paper 2 must show the directed / context layer carries information the shadow does not, measured as compression or held-out likelihood gain, not as a picture that "looks like" a law.
- G3 · Significance by permutation, done right. ≥ 1000 permutations; empirical $p = (k+1)/(B+1)$; report the null distribution. Never a σ from a handful of runs. Pre-specified nulls: language-label permutation, operator permutation, context shuffle, concept permutation.
- G4 · Discover-then-contrast with a *frozen* law inventory. The known-law set (Grimm, Verner, Grassmann, rhotacism, Polynesian lenition, …) is written down and hashed *before* inspecting components. We report all components/rules and the fraction that map to the inventory vs.

remain unexplained, with a defined matching metric (ARI/NMI, chance-adjusted) — no cherry-picked "this one is Grimm."

- G5 · No genealogy in, ever. Family membership only *delimits* the field; no tree, branch label, or protoform enters discovery. Branch labels are used only for *post-hoc* scoring, and because the branch cut is sample-dependent, always chance-adjusted and checked across cut granularities.
- G6 · Direction honesty. Only *documented dated stages* license an arrow of change. Distributional/tensor asymmetry is reported as suggestive, explicitly not causal; what stays undecidable is stated.
- G7 · "Law" is operational, not asserted. A context-conditioned bundle earns the name *rule/law* only by a stated MDL criterion — description-length (or held-out likelihood) gain over the best baseline code (G2), not over a straw-man. Rank/component count chosen by held-out fit or stability, with reconstruction error reported.
- G8 · Stability, not a single run. Bootstrap over languages *and* concepts; vary the seed; report component / rule stability (e.g. matched-component correlation, ARI across resamples), never one lucky factorization.
- G9 · Provenance & confounds from the start (not "future work"). Exact doculects, contexts, and datasets listed; leave-one-dataset-out, one-doculect-per-Glottocode, and source-blocked permutation run in the first results section, because dialect density and shared sources are known confounds (Paper 1 §7 lesson).
- G10 · Reproducibility & scope. Pinned deps, fixed seeds, config-keyed caches, bundled manifests/results from day one; irreducible stochasticity (e.g. LexStat's scorer) disclosed. Scope kept modest — one family with enough languages, well-defined contexts; no cross-family "how it diversified" claims (the H4 lesson — that question is reserved for the research programme with matched sampling).

Falsifier for the whole paper (stated up front): if the directed/context layer shows no compression or held-out-likelihood gain over the order-0 and baseline models (G2), and no component/rule survives the permutation nulls (G3) or maps to the frozen inventory above chance (G4), then the directed layer adds nothing and the paper reports that null result rather than dressing it up.

0.5 Research narrative — how this path was found (kept on the record)

This paper is deliberately written as a two-moment study; the arc between the moments is the contribution, so we record how it was found rather than presenting a polished afterthought.

Moment 1 — the make-or-break (**`paper2/context_test.py`**). Before building rule extraction we asked the single question that could kill the premise: *does the phonological environment predict a correspondence beyond the unconditioned within-pair correspondence?* We model the target segment b given (language pair, source segment a, environment), by held-out cross-entropy, split by cognate set, against a within-(pair,segment) environment-permutation null. Getting to a trustworthy number took three honest corrections — each a guardrail (§0) catching an artefact before we believed a result:

1. Scalability. The first version aligned all pairs with a pure-Python Needleman–Wunsch (capped at 25 languages). Replaced by lingpy multiple-sequence alignment (SCA, C-optimised): events read off aligned *columns*, cost independent of family size² — now scales to many languages.
2. A vacuous object. The first "operator" $\Delta(a, b)$ is a *deterministic* function of the segment pair, so conditioning on the segment already predicted it perfectly and left nothing for context. Re-

framed to the genuine sound-law object: predict the *target segment* from *source* + *environment*, directionally.

3. Broken estimation. Additive smoothing over-trusted singleton contexts; the sanity check ($P(b \mid$ pair, $a)$ should beat $P(b \mid a)$) *failed*. Replaced by hierarchical Witten–Bell interpolation; the sanity check then passed (+0.32 bits), validating the machinery.

What Moment 1 found (n = 25 IE, robust across coarse and rich environments): the environment is not noise — it beats a shuffled environment at $p = 0.002$ — yet it never pays for itself in average held-out compression (net ≈ -0.3 bits), even enriched from 21 to 103 environment classes. The reconciliation is the finding: context-conditioning of correspondences is real but *sparse*. Most correspondences are regular (unconditioned); only a few are context-conditioned (Verner-type splits). An average-over-all-correspondences metric washes the few real rules out; the permutation significance detects that they exist. Therefore the average-compression test is the wrong instrument, and Moment 2 must target the sparse rules directly (MDL).

A recoverability boundary (a small formal point). The corpus (Lexibank surface segmentation) carries no stress/prosody (0% of forms) and only inconsistent morpheme boundaries. Verner's law conditions on Proto-Indo-European accent — information *absent* from segment-only data — so no reconstruction-free method on this corpus could recover it; it is information-theoretically out of reach, not a failure of ours. This motivates a partition of sound laws by *what they condition on*, and a statement of which classes are recoverable from segment-only data and which are provably not. (The complementary vision — building a corpus that layers phonetics $\rightarrow$ prosody $\rightarrow$ polysemic semantics — is a separate, future infrastructure thread, not part of this paper.)

Moment 1b — the unicity distance of the key (**`paper2/unicity_test.py`**). The cryptographic instrument suggested a second, orthogonal measurement of the *same* per-pair key that 2A characterised: not *is there context?* but *how many cognates does it take to FIX a correspondence?* — the unicity distance $n^\star$ of §4 of the framing paper, which is strictly stronger than "beats a permutation null." For each key cell (language pair $\times$ source segment) we accumulate cognates one at a time (unit = cognate set, so no pseudo-replication) and mark it *fixed* when one target carries $\geq \theta$ of the posterior mass (weak prior of total mass 1, so the answer reflects the data, not the alphabet size), averaging over 20 seeded orderings. Result (§1.1): where a correspondence is regular it is fixed by a handful of cognates ($n^\star \approx 7$ at $\theta{=}0.7$) — the breakable-cipher result, now a number — and real cells concentrate far more than a marginal-shuffle null (493 vs 10), confirming genuine key signal. But only a minority of cells concentrate at all; the large censored majority upper-bounds genuine many-to-one (mergers) because it also absorbs cognate-detection noise. This refines 2A's "monoalphabetic" verdict: the per-pair key is *regular but lossy*, exactly the residual the 3.7-bit conditional entropy already implied.

Moment 2 — targeted rule extraction (next). Find the specific `(pair, source, env → target)` rules that individually beat their unconditioned baseline by a large margin, verify they are not sampling artefacts, and ask whether they read as real conditioned splits — then formalise with MDL and contrast (post hoc, frozen inventory) against known laws.

0.7 Theoretical framing (the cryptographic instrument)

Paper 2 adopts the programme's framing paper, *The Cipher of Sound Change* (`docs/cryptographic-framing.en.md`), and states its question in that vocabulary. A language pair is two ciphertexts in depth; the within-pair correspondence $P(b \mid \text{pair}, a)$ is the substitution key. Paper 2's single question is then: *is that cipher monoalphabetic (context-free) or polyalphabetic (its substitution alphabet switches with a context/keystream)?* The environment model $P(b \mid \text{pair}, a, \text{env})$ is the polyalphabetic hypothesis; the permutation null asks whether any polyalphabetic structure exceeds a shuffled keystream; and MDL asks whether a conditioned rule pays its key-cost (shortens the description of key + data). Two

consequences of the framing are load-bearing here: (i) a conditioning law whose key was overwritten in the ciphertext — Verner's Law, keyed on an accent later moved and erased — is *uncrackable in principle* from ciphertext lacking that key, which is exactly our corpus's limit (no stress); (ii) "is it a law?" is not a matter of taste but of description-length gain over the monoalphabetic key. The full development, taxonomy, and information-theoretic quantities live in the framing paper; this section only imports its terms. Two of the framing paper's refinements matter for how we read the results below: the key is properly a stochastic channel $K_{A\to B}(b \mid a)$, not a bijection (so mergers and probabilistic correspondence are native), and unicity distance is not a permutation test — it asks *how many cognates fix the key*, which §1.1 now measures directly. (The framing paper also opens a vowel-frequency channel and a wider mathematical-endolinguistics arm; those are not exercised here.)

1. Moment 1 — results and verdict (Indo-European, n = 25; current state)

In one line. Between related languages the sound correspondence is regular and context-free — a substitution cipher with a single alphabet, not one that switches with context.

From the seeded run of `paper2/context_test.py` (25 doculects, MSA over statistically-detected cognate sets, 417,696 directional correspondence events over 128 source segments; held-out cross-entropy 5-fold by cognate set, Witten–Bell hierarchical smoothing).

model (predicting the target segment b)	bits/event (held-out)
$P(b)$	4.60
$P(b \mid a)$ (source segment)	4.04
$P(b \mid \text{pair}, a)$ (within-pair correspondence)	3.73
$P(b \mid \text{pair}, a, \text{env})$ (+ environment)	4.02

- Sanity (machinery valid): the within-pair correspondence buys +0.32 bits over source-only — regular correspondences are captured.
- Decisive quantity: the environment buys −0.29 to −0.32 bits (it does *not* compress on average), yet it beats a within-(pair,segment) environment-permutation null at p ≈ 0.002. So the conditioning signal is real but not compressive on average → sparse.
- Moment 2, targeted extraction: the four strongest candidate rules were all word-final and phonologically implausible (ɾ→t, m→t). A word-internal control (strip initial/final positions, `POS_FILTER=m`) removed them entirely (4 → 0) while the diffuse environment signal persisted (p ≈ 0.002). Verdict: the strongest apparent conditioning is morphological / edge artefact (affixes aligning across languages at word edges), not phonology; word-internal, a weak diffuse phonological conditioning remains but yields no extractable, clean sound-law rules at this resolution.
- A recoverability boundary: stress is absent from the corpus, so Verner-class (accent-conditioned) laws are unrecoverable in principle — not a failure of the method.

The honest headline, in cipher terms. The sound-change map is essentially monoalphabetic — a context-free substitution, one key per language pair. $P(b \mid \text{pair}, a)$ dominates, and adding the environment ($P(b \mid \text{pair}, a, \text{env})$) does not help. It is not polyalphabetic.

The famous context-sensitive laws (Verner) are the polyalphabetic *exceptions*, and here they are rare or corpus-invisible. §1.1 then sharpens *what kind* of monoalphabetic key this is: regular but lossy. Where a correspondence does concentrate, about seven cognates fix it — but most cells stay many-to-one. That is a quantified, falsifiable claim about the nature of sound change.

1.1 Experiment 2B — the unicity distance of the per-pair key (n = 25)

From the seeded run of `paper2/unicity_test.py` over the same cached correspondence events; unit = cognate set; concentration = top-1 posterior mass $\geq \theta$ under a weak (total-mass-1) prior; means over 20 seeded orderings; 6,974 key cells (language pair × source segment) with ≥ 20 cognates. Persisted to `data/db/paper2.db` (`unicity`, `unicity_curve`, `unicity_cell`).

The unicity distance $n^\star_\theta = \min\{n : \text{top-1 mass} \geq \theta\}$ is the classical cryptanalytic question — *how much ciphertext (here: how many cognates) to fix the key?* — and it is strictly stronger than 2A's permutation test (which only asks whether structure beats chance).

criterion	REAL: cells that fix (of 6,974)	REAL median $n^\star$	NULL (marginal-shuffle): cells that fix
$\theta = 0.7$	493 (7 %)	≈ 7 cognates	10 (0.1 %)
$\theta = 0.9$	76 (1 %)	≈ 13 cognates	0

- Where the key is regular, it is *broken* by a handful of cognates — median $n^\star \approx 7$ at θ=0.7. The cleanest cells (intra-branch $r \to r$, $n \to n$, $l \to l$) fix in 3–7. This is the first empirical value of the framing paper's §4 unicity quantity.
- The concentration is genuine, not counting: real cells fix $\approx 49\times$ more often than a marginal-shuffle null (493 vs 10 at θ=0.7; the null reaches θ=0.9 for *zero* cells). Beating the null here means the correspondence carries pair-specific key structure beyond the source's marginal $P(b \mid a)$.
- But the per-pair key is lossy. Only a minority of cells ever concentrate a single target (mean dominant-target mass ≈ 0.34, flat in n); the large censored majority marks many-to-one correspondences. This refines — does not contradict — 2A: the "monoalphabetic" key is real yet carries the 3.7-bit residual entropy 2A measured, i.e. regular *and* lossy.
- Honest bound (G2/G9). The censored fraction is an upper bound on true non-identifiability: it also absorbs LexStat cognate-detection noise (false pairs → random targets). The next subsection tightens exactly this.

1.2 Gold-cognacy decomposition — how much of the censoring is *method*, not *language* (item D, done)

From `paper2/collect_iecor_events.py` + `paper2/unicity_test.py`: we hold the lexicon fixed to IE-CoR (iecor) and assign cognacy two ways — expert (curated cognate sets) and LexStat on the very same forms — plus the original Lebixank/LexStat run for reference. Table at $\theta = 0.7$, REAL vs the marginal-shuffle NULL; persisted to `data/db/paper2.db` (`cognacy` column).

cognacy source	lexicon	cells ($\geq$20 cog)	median $n^\star$	censored (REAL)	censored (NULL)
LexStat	Lexibank IE (25 doc)	6,974	7	93 %	100 %
expert (gold)	iecor IE (25 lang)	186	4	11 %	56 %
LexStat	iecor (same forms)	128	3	4 %	34 %

This revises 2B's reading, honestly and in our favour:

- The 93 % was overwhelmingly a corpus artefact, not structural non-identifiability. On curated IE data with expert cognacy, only 11 % of well-attested cells fail to concentrate — i.e. 89 % of per-pair keys *do* fix, at a median of 4 cognates. The bottleneck was the data (breadth + statistical-cognacy noise), exactly the hypothesis that motivated the corpus upgrade.
- The signal margin over chance widens sharply: REAL 11 % vs NULL 56 % censored (gold) — a far larger gap than Lexibank's 93 % vs 100 %. Cleaner cognacy makes the key *much* more clearly recoverable.
- A circularity caveat that makes gold the honest number. On identical iecor forms, LexStat censors *less* than expert cognacy (4 % vs 11 %). That is not LexStat being better — LexStat *defines* cognacy by phonetic similarity, so it fragments divergent-but-related forms out of a set and inflates within-cluster concentration (partly circular for a concentration measure). Expert cognacy keeps the genuinely divergent cognates, so its 11 % is the honest residual diffuseness (real mergers/divergence), and its median $n^\star = 4$ the honest unicity distance.
- Caveat (scope). Lexibank→iecor also changes breadth and cell count (6,974 → 186 well-attested cells), so the 93 %→11 % drop is dataset + attestation + cognacy combined; the *clean* isolation of cognacy-method is the gold-vs-LexStat pair on identical forms (the circularity point). Still n = 25; scaling is future.

Upshot. The per-pair substitution key is even more clearly regular than the Lexibank run implied — where attested, it is broken by 4 cognates — and its residual non-identifiability is small (11 %) and real. Better lexicons make the experiment sharply better, which is why the corpus upgrade preceded Paper 3.

1.3 Moment 2 — MDL rule extraction (IE-CoR gold, word-internal)

From `paper2/moment2_rules.py` on the gold event set, word-internal. A rule `(pair, source a, env → target)` is kept only if it compresses — the bits it saves by predicting its events under the conditioned distribution instead of the unconditioned $P(b \mid \text{pair}, a)$, minus the bits to state the rule (MDL, guardrail G7) — and flips the modal target.

On Lexibank the four strongest candidate rules were word-final and vanished under the word-internal control (§1). On clean gold cognacy, word-internal, the result is sharper.

- The aggregate signal is real but non-compressive, exactly as in Moment 1: environment still beats a shuffled-environment null ($p = 0.002$), yet costs -0.02 bits/event on held-out average.
- No rule survives MDL. Of the 26 `(pair × source × environment)` sub-cells with enough data, zero show a conditioned modal target different from the unconditioned one — the correspondence is the *same whatever the neighbourhood* — so there is nothing to extract; every candidate split *loses* bits even before the 19-bit rule-statement cost.

Verdict. With artefact-free data and a fair MDL criterion, the conditioned (polyalphabetic) layer yields no extractable rules at this resolution: the environment does not flip a single correspondence. This is the strongest form of the Moment-1 finding — to first order the cipher is monoalphabetic, and the real-but-faint conditioning is a distributional wobble, not a Verner-type split. *Honest caveat:* word-internal gold at $n = 25$ leaves only 26 testable sub-cells (the sparsity Moment 1 flagged), so "no rules" is partly a statement about data scale; the declared next step is to scale n, and — for accent-conditioned laws — a corpus with prosody, which no segment-only corpus can supply.

2. Future work (declared, not attempted here)

- (B) Richer context. Replace the neighbour *feature buckets* with the neighbour's *full segment identity*, word-internal only — may surface genuine phonological rules, at the risk of overfitting (to be controlled by the same held-out + permutation discipline).
- (C) Scale. Run at n $\geq$ 50 and beyond (the MSA pipeline already scales) for more events per `(pair, source)` context, which sparsity currently limits.
- (D) Tighten the unicity bound — *done* (§1.2). Expert (iecor) cognacy shows the Lexibank 93 % censoring was mostly a corpus artefact: on curated IE data 89 % of well-attested keys concentrate at median $n^\star \approx 4$. *Remaining:* scale beyond n = 25 (few cells clear $\geq$20 cognates at this size), split intra- vs cross-branch, and add expert cognacy for a second family.
- (Corpus) The recoverability ceiling. Accent/prosody and consistent morphology are absent from Lexibank; a corpus layering phonetics $\rightarrow$ prosody $\rightarrow$ polysemic semantics would raise the class of recoverable laws (a separate infrastructure thread).

Abstract ‹stub›

One paragraph. The pilot studied the *symmetric* repertoire of feature-difference operators and found its additive geometry to be representational, not (at the type level) genealogical — the genealogical signal living one level up, in the *distribution* of correspondences across languages, contexts, and direction. This paper builds that level: a directed, context-conditioned model in which recurrent bundles of correspondences and conditioned rules are *discovered*, not supplied, and only afterwards contrasted with known sound laws. State the headline once results exist (which laws emerge; how much genealogy the distribution carries; what direction can be recovered reconstruction-free). ‹fill after results›

1. Introduction — from a symmetric shadow to a directed network

- Recap the pilot in two sentences: operator $\Delta(a, b) = 1[\phi(a) \neq \phi(b)]$; repertoire O; the additive structure is representational; genealogy lives in the distribution (pilot Annex A). ‹stub›
- The gap this paper closes: a bare operator is general phonology; a *sound law* is a coordinated, directed, context-conditioned bundle over specific languages. Genealogy is a property of that joint structure. ‹stub›
- The stance is unchanged: discover first, contrast later. We do not insert Grimm, Grassmann, Verner, or Polynesian lenition; we test whether they *fall out* as compressive structure. ‹stub›
- Contributions (list once written): the order-1 tensor and its bundles; the order-2 conditioned rules via MDL; reconstruction-free direction; a historical-holdout evaluation. ‹stub›

2. Objects: orders 0, 1, 2, and the directed operator

- Order 0 (pilot): the symmetric repertoire O and $C(O)$ — the unoriented shadow. ‹stub›
- Order 1: the tensor X_{ijo} = normalized frequency of operator o between languages i, j. Captures *which languages* share *which operators* and how often. ‹stub›

- Order 2: X_{ijoc} — add context c (preceding/following segment, position, stress, morphological boundary). The conditional $P(o \mid c, \ell_i, \ell_j)$ is the object a sound law actually is. ‹stub›
- The directed operator: $T = (a, b, c, p, \ell_1, \ell_2, w)$ with signature $\sigma(T) = \Delta(a, b)$ but $T_{a\to b} \neq T_{b\to a}$. Define precisely; relate to X_{ijoc}. ‹stub›

3. Data and method

- Corpora: Lexibank (IPA-segmented), IE-CoR; dated/attested stages for the time signal (documented dates, not reconstructed ancestors). ‹stub›
- Building the tensors: coderivative sets (statistical, as in the pilot) → aligned correspondences tagged by language pair, position, and context → X_{ijo}, X_{ijoc}. ‹stub›
- Bundle discovery (order 1): non-negative tensor factorization $X \approx \sum_q \lambda_q\, a_q \otimes a_q \otimes b_q$ — each component = a set of languages (a_q) × a bundle of operators (b_q). Model selection for the number of components. ‹stub›
- Rule discovery (order 2): Minimum Description Length — a context-conditioned set of operators earns the name *rule* when it minimizes $L(\mathcal{R}) + L(D \mid \mathcal{R})$; define the code precisely. ‹stub›
- Direction without reconstruction: enumerate the admissible sources of an arrow that do *not* import a reconstructed ancestor — documented dates t; distributional/tensor asymmetry; markedness-neutral cues — and state honestly what direction can and cannot be recovered. ‹stub›
- Nulls & controls: the pilot's discipline extended to the tensor (concept-permuted D_R; language-grouping null; language-level bootstrap of components). ‹stub›

4. Results I — emergent bundles (order 1)

Pre-registration of Experiment 1 (write-before-run, per §0). *Object:* the order-1 tensor X_{ijo} over IE language pairs (family delimits the field, G5), built from the same coderivative alignments as Paper 1. *Missingness (G1):* cells for pairs below `MINSLOT` are masked, not zero-filled; factorization uses a masked loss; coverage is reported. *Method:* masked non-negative factorization; rank chosen by held-out reconstruction error (G7). *Baselines (G2):* (i) order-0 symmetric repertoire; (ii) per-language marginal operator profile; (iii) raw operator frequency. A component-set counts only if it improves held-out likelihood / compression over all three. *Null (G3):* language-label permutation ($B \geq 1000$), empirical $p = (k+1)/(B+1)$, for the statistic "component $\leftrightarrow$ branch agreement (chance-adjusted ARI)." *Contrast (G4):* the subgroup inventory (Glottolog branches) is frozen in advance; we report every component and the chance-adjusted ARI/NMI of the whole component set to branches — not selected matches. *Stability (G8):* language- and concept-bootstrap; matched-component correlation across resamples. *Confounds (G9):* leave-one-dataset-out and one-per-Glottocode repeats. *Falsifier:* components do not beat the three baselines, or do not exceed the permutation null. *(An early, pre-guardrail run of `src/tensor_pairs.py` — Romance/Germanic/Balto-Slavic components with no mask, no baseline, no null, and post-hoc branch reading — is superseded and not cited as a result.)*

- The masked-NMF components of X_{ijo}: report all components (language-set × operator-bundle) with coverage and rank-selection curve. ‹fill›
- Baseline comparison (G2): held-out likelihood / compression of the component model vs order-0, marginal, and raw-frequency baselines. ‹fill›

- Contrast (post hoc, G4): chance-adjusted ARI/NMI of the full component set to frozen Glottolog branches, with the permutation-null p; stability across resamples; leave-one-dataset-out and one-per-Glottocode. ‹fill›
- Honest read: which bundles carry information beyond the shadow, which are baseline-explained. ‹fill›

5. Results II — context-conditioned rules (order 2) and emergent laws

- $P(o \mid c, \ell_i, \ell_j)$ and the MDL rule set. ‹fill›
- The test: do known laws (Grimm, Grassmann, Verner, Polynesian lenition, rhotacism) emerge as compressive, context-conditioned bundles *without being supplied*? Success criterion and falsifier (no component/rule beats chance against the known-law inventory). ‹fill›
- Worked emergent rule(s) in real words. ‹fill›

6. Direction — the arrow of time without reconstruction

- What the directed layer recovers from documented time and from asymmetry; what stays undecidable. ‹fill›
- The repertoire as an *unoriented shadow*; how much orientation the data restore. ‹fill›

7. Historical-holdout evaluation

- Blind discovery (no laws, no trees, in some runs no family labels) $\rightarrow$ external contrast (laws, reconstructions, chronologies, contact). Metrics: ARI, NMI, law-recovery rate. Non-match as diagnosis. ‹fill›

8. Discussion

- Genealogy realized as distributional structure (the pilot's Annex A, now measured). ‹stub›
- Relation to phylogenetic-network work (Nelson-Sathi et al.; François linkages): we discover bundles rather than add borrowing edges to a tree. ‹stub›
- What "emergent law" means, and its limits. ‹stub›

9. Limitations and scope ‹stub›

Tensor sparsity and rank selection; context coding choices; direction is partial; still corpus/aligner-dependent; pilot families only unless W1 has scaled. Everything beyond stays in the research programme.

10. Reproducibility ‹stub›

New `make` targets (tensor build, NMF, MDL, direction, holdout); datasets and versions; seeds; durable results store (`data/results/`).

References ‹stub›

Reuse the pilot's [1]–[22]; add: non-negative tensor factorization; Minimum Description Length (Rissanen; Grünwald); sound-change typology for the *contrast* set (Kümmel); any tensor-linguistics precedents.