Additive Structure of Phonological Correspondences

A protoform-agnostic method for discovering mathematical patterns in documented linguistic systems

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

Abstract

Historical linguistics infers ancestral forms and then derives documented ones from them. We invert that order. Representing each attested sound correspondence not as a directed change but as the set of phonological features that differ between two aligned segments — a symmetric *operator* — a language family becomes a repertoire O of feature-difference vectors over $\mathbb{F}_2$, computed with no protoforms, no sound laws, and no externally imposed sound classes. (We say *protoform-agnostic* rather than "reconstruction-free": cognacy is still detected statistically with LexStat, which is itself a correspondence-based inference — see §3.) We study the geometry and additive structure of O for Indo-European (IE) and Austronesian (AN), drawn from Lexibank via statistical cognacy (LexStat) and feature-distance alignment (panphon). Four results. (i) Occupancy is not sparsity of the languages but vastness of the space: against $\approx 1.6\%$ of the algebraic span, the repertoire fills a large fraction of what is actually *available* — 0.29–0.63 (IE) and 0.27–0.48 (AN) depending on whether availability is measured alignment-independently or from the aligner — an order of magnitude denser than the span, not sparse. (ii) The repertoire is additively organized — combining two operators lands back in O far more often than chance, and this survives a hierarchy of six nested null models up to and including sampling from the corpus's own opportunity set ($Z = +6.3$ IE, $+5.9$ AN; $p_{\mathrm{MC}} \leq 0.004$) — robustly in IE but only fragilely in AN under a language-level bootstrap (CI_{95} $[0.18, 0.22]$ vs $[0.15, 0.30]$). (iii) That additive structure is a property of the phonological representation, not of genealogy: a concept-permuted control matches it, and branch repertoires overlap no more than random language groupings in IE (whereas in AN they are significantly *more differentiated* than random — a genealogical signal). (iv) The generable-but-unattested operators $\langle O \rangle \setminus O$ form a concrete geometry of near-misses. The contribution is not a sound law but an *object of study* — the empirical repertoire of correspondences and its measurable additive geometry — together with an inference order that keeps the historical apparatus out of the computation so it can serve as an independent test. A companion manual gives the full didactic treatment.

1. Introduction

The comparative method is an inference toward a hidden past: from documented forms X it reconstructs an ancestor $\widehat{Z}$ and posits the changes H that produced X. It answers *how did these forms descend from their common ancestor?* We ask a different question — *treated purely as a set of differences, what shape does the space of attested correspondences have?* — and we answer it in the opposite order:

$$X \longrightarrow \mathcal{M}(X), \qquad \text{then, separately,} \qquad \mathcal{M}(X) \longleftrightarrow H,$$

where $\mathcal{M}(X)$ is a mathematical structure discovered without protoforms or precoded laws, and the historical apparatus enters only *afterward*, as contrast. This is not a rejection of reconstruction; it is a more demanding test of it. Protoforms are inferred *from* the documented languages; building

the space *on* them and then checking whether it confirms historical regularities risks circularity — leakage of the target into the inputs. Suspending them during discovery is therefore a design choice, not a limitation.

Two clarifications frame everything below. First, the method is not "theory-free": transcription, segmentation, concept assignment, the feature matrix, and the aligner are measurement instruments whose influence we make explicit and control for (§3, §4.3). Second, this is emphatically not a claim that "sound change is XOR," nor a hunt for known sound laws; it is a change of what is measured and in what order.

Two further commitments follow, and we state them plainly for the comparative linguist. First, sound laws are precise about *that* a change occurred and *what* resulted, but silent about *why* one related system selects one outcome where another selects a different one — a "why" that may turn on substrate, contact, and areal pressure, because a language is less a branch off one parent than a new formation confluent from many sources (substrate included as *constitutive*, not noise); such questions live on a network rather than a tree, and become *measurable* only once one can see the whole space of what a system *could* have done (the "holes" of §4.4). Second, and for the same reason, we call two related forms coderivatives rather than *cognates*: the family metaphor (sister, mother, daughter tongues) presupposes descent from a single ancestor, which is exactly the commitment we withhold from the computation. A coderivative is a form co-derived with another, its relatedness evidenced by systematic correspondence; a cognate is the special case of clean descent from one ancestor. None of this denies proto-languages or faults reconstruction — it adds a mathematical dimension alongside the symbolic one, and renames only to keep the question open.

1.1 Related work

This work sits inside computational historical linguistics but asks an unusual question of it. Data and tools. We build on Lexibank [1], the aggregated, IPA-standardized wordlist repository; panphon [2] for articulatory feature vectors; LingPy/LexStat [3] for statistical cognate detection; Glottolog [4] for genealogical classification; Concepticon [5] for comparison meanings; and IE-CoR [6] for expert Indo-European cognacy used only in validation (§4.6). Alignment uses Needleman–Wunsch [7]; feature-based phonetic alignment and cognate scoring descend from Kondrak [8] and Dolgopolsky sound classes [9].

Where the field concentrates, and where we differ. Most computational work uses sound correspondences *instrumentally* — as a means to detect cognates, build phylogenies, or date splits: LexStat [3], global-scale phylogenetic inference [10], and cognate-detection benchmarks [11] all treat the correspondence as a step toward a *tree* or an *ancestor*. Typological work such as Blasi et al. [12] tests sound–meaning associations across thousands of languages, again with the correspondence/segment as an instrument. Our object is different: we study the set of feature-difference operators itself — its geometry, additive structure, and occupancy — as the primary phenomenon, and we deliberately withhold the historical apparatus (protoforms, laws, sound classes) from the computation so it can serve as an *independent* test rather than an input. Feature-difference encodings and sound-change typologies exist (e.g. Kümmel's survey of consonant change [13]); what is new here is not the encoding but treating the resulting *repertoire* as a measurable region of a feature space and characterizing it with explicit null models. Confluence, as support. Our stance that a language is a confluence rather than a branch descends from a long tree-skeptical line — Schmidt's wave model [17], Schuchardt's insistence on mixture [18], Trubetzkoy's Sprachbund [19], and modern phylogenetic networks and linkages [20–22]; we invoke it as *support*, but where those add horizontal edges to a framework of descent (borrowing as deviation from a tree), we withhold descent from discovery and study the geometry of coderivation directly. Mathematics. The additive/geometric apparatus draws on additive combinatorics [14] (sumsets, doubling, additive energy), matroid theory [15] (the dependency structure of the operator set), and coding theory [16] (the span as a linear code, the geometry of unrealized

"holes"). To our knowledge these have not previously been applied to phonological correspondence repertoires.

2. The operator and the repertoire

For two aligned segments a, b with panphon feature map ϕ, the operator is the indicator of differing features

$$\Delta(a, b) = \mathbf{1}[\phi(a) \neq \phi(b)] = \{k : \phi_k(a) \neq \phi_k(b)\} \in \mathbb{F}_2^n.$$

This is a well-defined $\mathbb{F}_2$ vector for *any* feature alphabet. Over strictly binary features it coincides with $\phi(a) \oplus \phi(b)$; but panphon features are ternary $(+1, 0, -1)$, so treating composition as XOR is a *binary approximation* — an identity on binary sub-features whose $\approx 2\%$ failures (in IE) precisely localize the non-binary dimensions (height), a diagnostic we return to. Δ is a *contrast*, not a directed change: $\Delta(a, b) = \Delta(b, a)$, so Δ does not distinguish $t \to s$ from $s \to t$, and "each operator is its own inverse" is a consequence of that symmetry, not a phonological finding. Directed history, when wanted, requires a richer object $T = (a, b, c, p, \ell_1, \ell_2, w)$ with signature $\sigma(T) = \Delta(a, b)$; this paper studies the symmetric signatures. We keep operators of one to three features (atoms and small molecules).

The forms compared are coderivatives (§1): related by a shared, evidenced history — recurrent correspondence — without presupposing a single reconstructed ancestor. A family's repertoire O is its set of recurrent operators over those coderivatives. O generates a subspace $\langle O \rangle = \text{span}_{\mathbb{F}_2}(O)$ of dimension (rank) r; crucially O is not a subspace (not closed under XOR, does not contain 0) — it is a *region*, and the object of interest is the relation between the region O and the capacity $\langle O \rangle$.

3. Data and method

Corpus. Lexibank word lists, IPA-segmented, for Indo-European (304 languages available) and Austronesian (978), taking per family the languages with the most forms. Cognacy is detected statistically with LexStat (LingPy); alignment is Needleman–Wunsch with substitution cost equal to the fraction of panphon features that differ; features are a 12-feature primary subset of panphon — `cont, voi, nas, ant, cor, lab, back, round, strid, hi, lo, son` (manner, place, laryngeal, and major-class features; panphon's full set is larger, and the subselection is a documented, auditable choice, swept in the sensitivity analysis). No protoforms, laws, or sound classes enter the computation. Operators are taken from aligned, non-identical, consonantal positions with support thresholds (30 at family scale); vowel correspondences are set aside in this pilot (they are dominated by ablaut/reduction in IE and would need a graded treatment — a deliberate scope limit, not a claim that they are noiseless). Independent validation uses IE-CoR expert cognacy (§4.6). Every quantity is reproducible from a single `make` target (§7).

On "protoform-agnostic." We withhold protoforms, sound laws, and sound classes from the discovery computation. We do *not* claim to be free of all historical inference: statistical cognacy (LexStat) learns language-pair sound correspondences and clusters by them, so a genealogical/correspondence assumption is present in how cognate sets are formed. But that assumption is the Neogrammarian one — cognacy evidenced by *recurrent, systematic correspondence* (what LexStat measures), which is the evidence *for* cognation, not the reconstructed ancestor inferred on top of it. We keep this foundation and set aside only the reconstructive step. Filtering the input to reconstruction-certified cognates would defeat the design — it would leak the historical answer into the data and forfeit the independent test — so what is kept out is the *reconstructed ancestor and the pre-stated laws*; §4.6 quantifies how much LexStat's choices shape the repertoire, and §4.3 controls for representation. The honest scope is therefore "protoform- and sound-law-free," not "assumption-free."

Two families give a minimal comparative frame; "IE" and "AN" are pilot genealogical boundaries, not results.

4. Results

4.1 Occupancy: sparsity is mostly algebraic

IE uses $|O| = 67$ operators, AN $|O| = 22$, with ranks $r = 12$ and $r = 10$. Against the full algebraic span these are 1.6% and 2.2% — but the span $2^r - 1$ contains feature-bundles that no pair of real segments could realize. Nested between O and $\langle O \rangle$ are the realizable universe $U_S = \{\Delta(a, b) : a, b \in S\}$ (over the attested segment inventory S) and the opportunity universe Ω_D (differences that actually occur in aligned corpus positions):

family	ρ_{alg} (span)	ρ_{seg} (realizable U_S)	ρ_{opp} (aligner Ω_D)	ρ_{ind} (alignment-fr
Indo-European	0.016	0.48	0.63	0.29
Austronesian	0.022	0.43	0.48	0.27

where $\rho_{\text{alg}} = |O|/(2^r - 1)$, $\rho_{\text{seg}} = |O|/|U_S \cap \langle O \rangle|$, and $\rho_{\text{opp}} = |O|/|\Omega_D \cap \langle O \rangle|$. Because Ω_D is built from our *own* alignments, ρ_{opp} could be circular; we therefore add an alignment-free opportunity Ω_{bag} (differences between consonants co-present in same-concept words, no alignment used) giving ρ_{ind}. The aligner did inflate the figure — ρ_{ind} is roughly half of ρ_{opp} — but the conclusion holds under a denominator that never touches our aligner: at 0.27–0.29 (and $\rho_{\text{seg}} \approx 0.43$–$0.48$, also alignment-free) the repertoire is an order of magnitude denser than the span. The "98% unused" is the algebraic space being enormous, not languages being restrictive.

4.2 The repertoire is additively organized

Composing two operators is XOR of their feature-sets. The composition-realization index

$$C(O) = \frac{|\{\{u, v\} \subseteq O : u \oplus v \in O\}|}{\binom{|O|}{2}}$$

is 0.220 (IE) and 0.234 (AN); conditioned on the opportunity universe, $C_\Omega(O) = 0.81$ (IE), 0.66 (AN), and conditioned on the alignment-free opportunity Ω_{bag} (§4.1) it remains substantial at 0.57 (IE), 0.48 (AN) — the composition-closure is not an aligner artifact. To test whether this exceeds chance we compare against a hierarchy of null repertoires, each preserving more structure and each sampled to size $|O|$; we report Z and empirical p_{MC} over 500 simulations:

null preserves	IE Z	AN Z
size + weight ≤ 3	+15.7	+12.0
Hamming weights	+14.8	+9.0
exact margins (swap-MCMC)	+7.5	+3.6
rank / span $\langle O \rangle$	+45.6	+13.7
realizable U_S	+9.2	+7.3
opportunity Ω_D	+6.3	+5.9

All six survive at $p_{\text{MC}} \leq 0.004$ (500 simulations); since every p_{MC} sits at the resolution floor, a Benjamini–Hochberg correction across the null $\times$ family $\times$ statistic grid leaves them all significant. (A larger B for finer p-values is deferred to the research programme.) We describe the six as "increasingly constrained" rather than a strict nesting: the span null (row 3) samples a different ambient set

than the others, so the Z column is not monotone. Even drawn from exactly the differences the corpus made available, the observed repertoire is significantly more composition-closed. Three further additive statistics agree (against the Ω_D null), each on unordered tuples to match $C(O)$: triple density $\tau(O) = |\{\{u, v, w\} \subseteq O : u \oplus v = w\}|/\binom{|O|}{3}$ is high (IE $Z{=}{+}6.6$, AN $+6.5$); the doubling constant $\kappa = |O \oplus O|/|O|$ (with $O \oplus O = \{u \oplus v\}$) is low (IE $Z{=}{-}5.1$, AN -6.0; small doubling = strong additive structure); and the additive energy $E(O) = |\{(a, b, c, d) \in O^4 : a \oplus b = c \oplus d\}|/|O|^3$ is high (IE $+7.4$, AN $+7.3$). The repertoire behaves like an approximate (possibly affine) subspace. Its dependency structure has a basis-independent core: the vector matroid of O (operators as $\mathbb{F}_2$ vectors) has 162 (IE) and 18 (AN) three-element circuits (minimal XOR-zero triples), with `{voi}` (voicing) in the most of them — the additive hub. *(Circuit and hole counts scale with $|O|$, so IE-vs-AN raw counts are not directly comparable; the size-controlled comparison is the Z against nulls.)*

Robustness (language-level bootstrap). Because $C(O)$ is a U-statistic over non-independent operator pairs, the valid resampling unit is the language: we resample languages with replacement and rerun the whole pipeline ($B = 40$). The result separates the two families sharply — $C(O)$ $\mathrm{CI}_{95} = [0.18, 0.22]$ for IE (tight, comfortably above the null) but $[0.15, 0.30]$ for AN (wide, lower end near the null). IE's additive structure is robust to which languages are sampled; Austronesian's is fragile — a caution the size-matched nulls did not surface, and a reminder that with a small, variable repertoire the per-family sample matters. (The language bootstrap was run for $C(O)$; the companion statistics τ, κ, E rest on the size-matched Ω_D null, and their language-level resampling is deferred.)

4.3 What the additive structure is — and is not

The structure of §4.2 is a property of the phonological representation and inventory, not of genealogy or semantics. Two controls establish this. First, permuting concept labels to destroy all cognacy and semantic linkage (regime D_R) yields additive structure *as high as or higher than* expert cognacy D_G ($C{=}0.215$ vs 0.169), so the type-level additive closure does not distinguish history from inventory. Second, the one linear dependency we observe — in AN, `ant` $=$ `cor`$\oplus$`lab` — is a property of the *selected* repertoire, not of the feature matrix: the realizable universe U_S has full rank (12, corank 0), so the ontology would permit independence; the dependency appears only in O_{AN}. The additive geometry is real (§4.2) but must be read as *representational*; genealogical signal lives elsewhere (§4.5–4.6).

4.4 The geometry of holes $\langle O \rangle \setminus O$

Among the operator-shaped (weight ≤ 3) vectors the span generates, most are holes — generable but unattested. For each hole x, its depth is $d(x, O) = \min_{o \in O} d_H(x, o)$.

	generable (weight ≤ 3)	observed	holes	at $d = 1$	holes in Ω_D
Indo-European	298	67	231	160	40
Austronesian	119	22	97	53	24

Most holes lie one feature from an attested operator; the linguistically loaded ones are the holes in Ω_D (40 IE, 24 AN) — differences the corpus made available yet the system left unused. These are the concrete candidates for interpretation: absences that are choices, not impossibilities.

4.5 Systems and branches

Partitioning each family into Glottolog branches and taking $O_F = \bigcup_r O_r$, the persistence histogram H_k (operators in exactly k branches) decomposes the repertoire without double-counting: IE $|O_F| = 97$, strict nucleus $K_\cap = 12$ (the atoms `{ant}` `{cont}` `{voi}` in every branch), majority nucleus

$K_{1/2} = 38$; AN $|O_F| = 48$, $K_\cap = 4$, $K_{1/2} = 18$. But raw branch overlap is uninterpretable without a null. Against a grouping null (arbitrary same-size language groups) the mean Jaccard tells opposite stories in the two families:

	mean Jaccard (real branches)	arbitrary groupings	Z
Indo-European	0.42	0.45 ± 0.05	-0.7 (indistinguishable)
Austronesian	0.51	0.57 ± 0.01	-5.4 (real branches more differentiated)

At the level of operator types, Indo-European branches share no more than random groupings — the apparent "shared core" is shared *inventory* — whereas Austronesian branches are significantly more differentiated than chance. Under rarefaction the nucleus size is sample-sensitive, so branch *richness* is not interpreted; the robust facts are the histogram decomposition and the grouping-null contrast.

4.6 Validation against expert cognacy

Statistical cognacy could contaminate the repertoire with false cognates. Against IE-CoR expert cognacy the contamination is real at the instance level (pair precision 0.73, recall 0.33) but absorbed at the type level: LexStat's operator inventory has precision ≈ 1.0 and recall 0.66 against the expert repertoire — a false cognate almost always produces an operator already present. The type inventory is thus more robust than individual cognate identification. Where genealogical signal does live is the distributions: mutual information between operator and branch is modest but non-zero and larger in AN than IE (0.22 vs 0.14 normalized), consistent with §4.5; a few operators dominate each family (effective count $N_{\text{eff}} = 24$ of 119 IE, 18 of 63 AN — where 119/63 are the *instance-level* type inventories before the support threshold that defines the repertoire O of 67/22, distinct again from the branch-union O_F of 97/48 in §4.5).

4.7 Sensitivity: the headlines are measurement-invariant

The paper flags transcription, segmentation, the feature matrix, and the thresholds as *measurement instruments* (§1, §3); a sensitivity sweep confirms the headline signs do not depend on them. Sweeping the support threshold (15–60), IE's rank stays 12 and $C(O)$ stays 0.215–0.224 (AN: rank 10–12, $C(O)$ 0.17–0.23); sweeping the operator weight cap ($\leq 2, 3, 4$) and the feature subset (the full 12, minus `lab`, minus `{round,lo}`, and a manner+voice+place 8-set) leaves $C(O)$ in 0.21–0.37 (IE) and 0.14–0.28 (AN) — in every case well above the size-matched null floor (≈ 0.12–0.15, §4.2). Occupancy ρ_{opp} stays 0.5–0.7 (IE) and ~ 0.4–0.5 (AN) across thresholds. The one soft spot, reported honestly, is Austronesian at weight cap ≤ 2 ($C(O) = 0.14$, $|O| = 12$), where the repertoire is simply too small to resolve. Reproducible: `make sensitivity FAMILY="X"`. (The LexStat-threshold and aligner sweeps, which require the corpora, are scheduled.)

5. Figures

Two graph types make the repertoire visible; each figure below carries a self-contained caption. Figures 1–2 are the *segmental correspondence graph* G_S for Indo-European and Austronesian: nodes are segments, edges are recurrent correspondences coloured by change type. Figure 3 is the *operator graph* G_O for Indo-European: nodes are operators and an edge joins u, v whenever their composition $u \oplus v$ is itself an operator of the repertoire — so G_O is the visual face of the composition index

$C(O)$ of §4.2. (The two graphs, together with a *language graph* G_L over operator distributions, are the three-graph architecture the companion manual develops.)

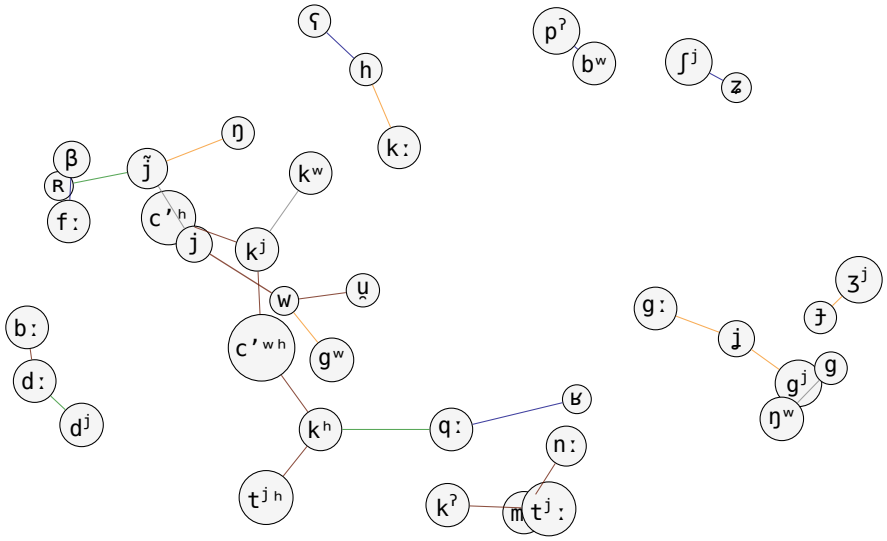

Figure 1: Segmental correspondence graph G_S of Indo-European. Nodes are segments; edges are recurrent correspondences, with thickness $\propto$ frequency and colour by change type: voicing, frication/lenition, palatalization/height, place, prenasalization.

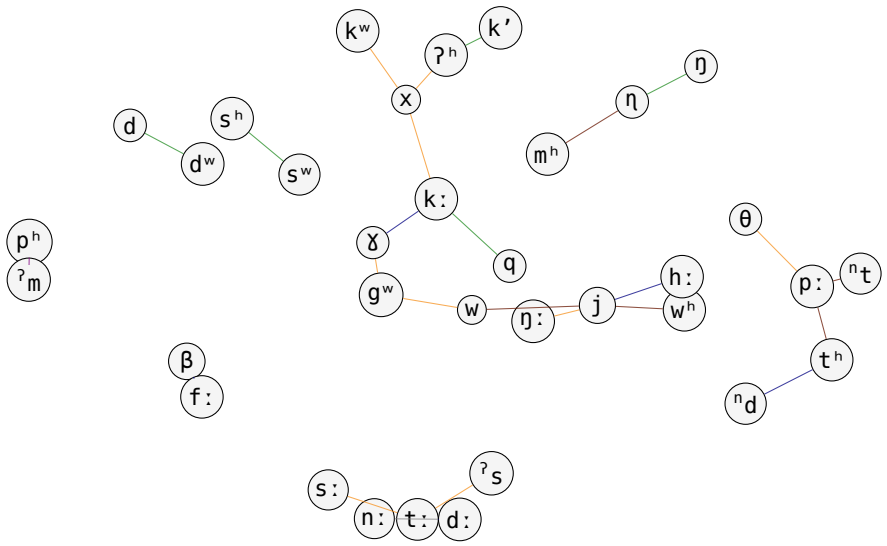

Figure 2: Segmental correspondence graph G_S of Austronesian. Nodes are segments; edges are recurrent correspondences, with thickness $\propto$ frequency and colour by change type: voicing, frication/lenition, palatalization/height, place, prenasalization.

6. Discussion

The contribution is an object and an order. The object is the empirical repertoire O of symmetric feature-difference operators, and the separation between it and the subspace $\langle O \rangle$ it generates: a system uses a small, additively-structured region of a large algebraic space, densely relative to opportunity, with a concrete geometry of unused near-misses. The order is discover-then-contrast, which keeps reconstructions, laws, and sound classes out of the measurement so each can serve as an independent yardstick later.

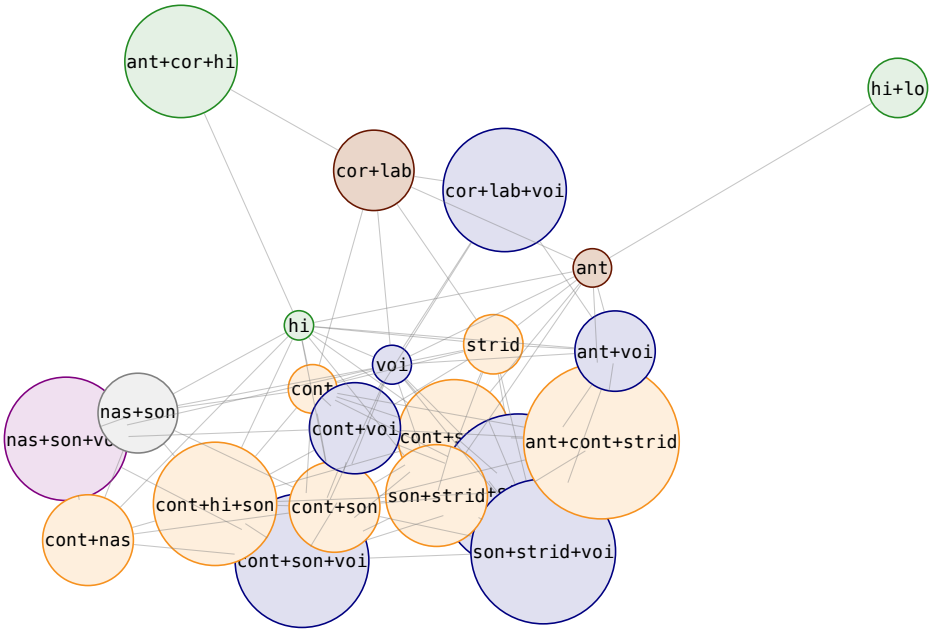

Figure 3: Operator graph G_O of Indo-European. Nodes are operators (the 22 with highest support), sized by frequency and coloured by change type (voicing, fricativization, palatalization/height, place, prenasalization). An edge joins u, v when their composition $u \oplus v$ is itself an operator of the repertoire — so this is the graph of *additive closure*. Dense connectivity is the visual face of the composition index $C(O)$.

Two boundaries keep the claims honest. The additive structure is *representational*, not genealogical (§4.3); the genealogical question is properly posed on distributions and, in future work, on a directed, context-conditioned layer — the tensor X_{ijoc} of operator frequency by language pair and context, whose non-negative factorization and minimum-description-length rules would let sound laws *emerge* as compressive bundles rather than be inserted. Under a historical-holdout protocol (discover; test stability; cluster languages blind; only then compare to reconstructions and trees), a recovered law would be evidence precisely because it was never an input. The same discipline — discover structure first, admit ground-truth labels only to test — is a template for leakage-free evaluation of language models on historical and typological tasks.

7. Limitations and reproducibility

Limitations. Two families are a pilot; ”family” is a genealogical label taken as a boundary, not a result. The opportunity universe Ω_D and the additive nulls depend on the corpus and aligner. The additive structure is type-level and representational; loans-as-system-influence could not be tested (loan annotations are absent in the pilot corpora and require a resource such as WOLD). Metathesis — a *reordering* of segments, as in the classic TR↔RT — is not captured: the alignment is monotonic (order-preserving), so a metathesized pair does not yield a ”metathesis operator” but surfaces as spurious substitutions and gaps (largely filtered by the alignment-cost cohesion screen, but a genuine gap). Metathesis is properly a *permutation* of skeleton positions, an operator on order rather than on the features of a slot — an orthogonal dimension (the symmetric group S_k beside the feature space $\mathbb{F}_2^n$) that a future model should add explicitly (see the research programme). The weight-≤ 3 operator cap excludes, by construction, compositions of weight > 3 from O while retaining them in the $C(O)$ denominator, deflating $C(O)$ somewhat; the sensitivity sweep (§4.7) shows the sign is invariant to the cap. Force-directed figures lack an articulatory coordinate system. None of these affect the null-controlled results, but they bound their interpretation.

Reproducibility. All results run from `transformations/src` against a repertoire built by `make family FAMILY="X"`. Key targets: `make universes` (§4.1), `make nulls` / `make additive` (§4.2, §4.4), `make repr-control` (§4.3), `make superposition` (§4.5), `make cognate-eval` / `make distributions` (§4.6). Datasets: Lexibank, IE-CoR, Glottolog (cached classification). LexStat threshold 0.55, 100-run scorer; panphon 12-feature primary set; seeds fixed. The companion manual (`docs/manual`) gives worked, hand-computed examples of every measure and the full derivations.

8. Scope and future work

This paper is a bounded pilot. Its claim is deliberately narrow: on two documented families, the empirical repertoire of feature-difference operators occupies a sparse-in-the-algebra but dense-in-opportunity, additively-structured region, and that structure is representational rather than (at the type level) genealogical — established with explicit null models, sensitivity sweeps, a representation control, a two-level cognacy validation, and a language-level robustness check. We claim no more than that, and we mark where even this is fragile (Austronesian's additive result).

Everything else belongs to the research programme (`docs/RESEARCH-PROGRAM.md`), not to this paper, and is explicitly out of scope here:

- Inferential comparison across families — moving from $n = 2$ to 20–50 families (W1).
- Where sound laws live — the directed, context-conditioned tensor and MDL-emergent rules (W2).
- The genealogical signal in distributions, and family-blind typology (W3, W6).
- Contact and time — loans-as-system-influence via WOLD; documented-date dynamics (W4, W5).
- Deeper mathematics — affineness tests, matroid/circuit and coding-theoretic invariants cross-family, persistent topology and spectra, and the order/permutation dimension of metathesis (S_k beside $\mathbb{F}_2^n$) (W7).
- Representation upgrades — the typed (ternary) feature space; an external, non-corpus opportunity universe; Austronesian gold cognacy (ABVD); LexStat-threshold and aligner sweeps; BCa/jackknife intervals and language-level resampling of τ, κ, E.

Drawing this boundary is the point: the object and the method introduced here are meant to *open* those studies, each its own paper, not to pre-empt them.

References

[1] List, J.-M., Forkel, R., Greenhill, S. J., Rzymski, C., Englisch, J., & Gray, R. D. (2022). Lexibank, a public repository of standardized wordlists with computed phonological and lexical features. *Scientific Data*, 9, 316.

[2] Mortensen, D. R., Littell, P., Bharadwaj, A., Goyal, K., Dyer, C., & Levin, L. (2016). PanPhon: A Resource for Mapping IPA Segments to Articulatory Feature Vectors. In *Proceedings of COLING 2016*, 3475–3484.

[3] List, J.-M. (2012). LexStat: Automatic detection of cognates in multilingual wordlists. In *Proceedings of the EACL 2012 Joint Workshop of LINGVIS & UNCLH*, 117–125. (See also the LingPy library: List, J.-M. & Forkel, R.)

[4] Hammarström, H., Forkel, R., Haspelmath, M., & Bank, S. (2023). *Glottolog* (version 4.x). Leipzig: Max Planck Institute for Evolutionary Anthropology. https://glottolog.org

[5] List, J.-M., Rzymski, C., Greenhill, S., Schweikhard, N., et al. (eds.) *Concepticon: A Resource for the Linking of Concept Lists.* https://concepticon.clld.org

[6] Heggarty, P., Anderson, C., Scarborough, M., et al. (2023). Language trees with sampled ancestors support a hybrid model for the origin of Indo-European languages. *Science*, 381, eabg0818. (Dataset: IE-CoR.)

[7] Needleman, S. B., & Wunsch, C. D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, 48(3), 443–453.

[8] Kondrak, G. (2000). A New Algorithm for the Alignment of Phonetic Sequences. In *Proceedings of NAACL 2000*, 288–295.

[9] Dolgopolsky, A. B. (1986). A probabilistic hypothesis concerning the oldest relationships among the language families of northern Eurasia. In *Typology, Relationship and Time* (Shevoroshkin, ed.).

[10] Jäger, G. (2018). Global-scale phylogenetic linguistic inference from lexical resources. *Scientific Data*, 5, 180189.

[11] Rama, T., List, J.-M., Wahle, J., & Jäger, G. (2018). Are Automatic Methods for Cognate Detection Good Enough for Phylogenetic Reconstruction in Historical Linguistics? In *Proceedings of NAACL-HLT 2018*, 393–400.

[12] Blasi, D. E., Wichmann, S., Hammarström, H., Stadler, P. F., & Christiansen, M. H. (2016). Sound–meaning association biases evidenced across thousands of languages. *PNAS*, 113(39), 10818–10823.

[13] Kümmel, M. J. (2007). *Konsonantenwandel: Bausteine zu einer Typologie des Lautwandels.* Wiesbaden: Reichert.

[14] Tao, T., & Vu, V. H. (2006). *Additive Combinatorics.* Cambridge University Press.

[15] Oxley, J. (2011). *Matroid Theory* (2nd ed.). Oxford University Press.

[16] MacWilliams, F. J., & Sloane, N. J. A. (1977). *The Theory of Error-Correcting Codes.* North-Holland.

Note (pilot preprint): bibliographic details verified against publisher records where possible; page numbers for a few workshop items should be confirmed at typesetting.

[17] Schmidt, J. (1872). *Die Verwandtschaftsverhältnisse der indogermanischen Sprachen.* Weimar: Böhlau. (The wave model, *Wellentheorie*.)

[18] Schuchardt, H. (1885). *Über die Lautgesetze: Gegen die Junggrammatiker.* Berlin: Oppenheim. (And his work on language mixture / *Sprachmischung*.)

[19] Trubetzkoy, N. S. (1928). Proposition 16. In *Actes du premier congrès international de linguistes*, The Hague. (The *Sprachbund* concept.)

[20] François, A. (2014). Trees, Waves and Linkages: Models of Language Diversification. In C. Bowern & B. Evans (eds.), *The Routledge Handbook of Historical Linguistics*, 161–189. London: Routledge.

[21] Nelson-Sathi, S., List, J.-M., Geisler, H., Fangerau, H., Gray, R. D., Martin, W., & Dagan, T. (2011). Networks uncover hidden lexical borrowing in Indo-European language evolution. *Proceedings of the Royal Society B*, 278(1713), 1794–1803.

[22] List, J.-M., Nelson-Sathi, S., Geisler, H., & Martin, W. (2014). Networks of lexical borrowing and lateral gene transfer in language and genome evolution. *BioEssays*, 36(2), 141–150.