Features in Typological Databases Are Multiple Choice, Not Yes/No Questions: Lessons From Feature-Value Embeddings

Anonymous ACL submission

Abstract

Typological NLP often turns a multi-valued database feature into independent binary targets. This is convenient for prediction, but it treats alternatives such as SOV, SVO, and VSO as unrelated yes/no labels. We compare two ways to link WALS and GRAMBANK values: direct profile matching over shared languages and learned value embeddings. For exact cross-database lookup in our benchmark, observed profiles are the best starting point: they are auditable, outperform learned vectors, and expose a strong genealogy confound. The embeddings are useful for a different question: how the values of one feature relate to one another. When the model treats those values as alternatives in one categorical choice, its vectors recover structure that independent binary models often miss, including ordinal scales, subtype trees, Boolean-like lattices, and cross-product contrasts. Controlled prediction and dependency-linearization checks provide targeted evidence that this structure matters beyond visualization. We recommend a simple workflow: use observed profiles to propose cross-database links, and use categorical embeddings to inspect relations among values of the same feature.

1 Introduction

Computational typology often represents multi-valued features with independent value indicators. If WALS 81A has the values SOV, SVO, VSO, VOS, OVS, OSV, and "no dominant order", a binary expansion asks seven separate questions: is the language SOV, is it SVO, and so on. This representation is familiar from typological feature prediction (Bjerva et al., 2019c), but it loses part of the structure of multi-valued features. SOV and OSV share verb-finality; SOV and SVO share subject-initiality; VSO and VOS share verb-initiality. A binary model may learn these relations indirectly, but nothing in the objective tells it that SOV and OVS share verb-finality, or that VSO and VOS share verb-initiality. Figure 1 illustrates this failure mode: the independent-binary baseline can invert an ordinal scale, merge distinct article values, and lose one axis of a cross-product contrast, while categorical softmax training more closely matches the annotated structure.

We instead model the values of a multi-valued feature as alternatives in one categorical choice. Genuinely binary features, common in GRAMBANK, should remain two-way contrasts. Many WALS-style features instead ask one question with several possible answers. A categorical model lets those answers compete, learns an embedding for each value, and makes the distances among values inspectable.

The same issue appears when databases are compared. WALS and GRAMBANK are central resources for typology-aware NLP, but they are not two encodings of the same schema (Haspelmath et al., 2005; Dryer and Haspelmath, 2013; Skirgård et al., 2023; Baylor et al., 2023). WALS often records a dominant value of a broad feature; GRAMBANK often asks whether a more specific property is present, absent, possible, or constructionally realized. Some WALS values have near-direct GRAMBANK counterparts. Others correspond to several GRAMBANK properties, imply a property without being equivalent to it, or simply co-occur because of genealogy, area, diachrony, or sampling.

The linguistic motivation is that typological features are comparative concepts. They are built for cross-linguistic comparison; they are not simply language-internal categories copied from grammars (Haspelmath, 2010; Croft, 2003). That position is not uncontested; cross-linguistic formal-category approaches remain an important counterpoint (Newmeyer, 2007). It is also compatible with autotypologizing approaches, where the set of types is revised as new languages and construc-

tions enter the comparison (Bickel and Nichols, 2002). We do not claim that grammar itself is discrete. Token-based, gradient, and measurement-oriented typology often represent word-order preferences and constructional distributions as frequencies, entropies, or continuous measurements over use (Bickel, 2007; Baylor et al., 2024; Levshina, 2019; Levshina et al., 2023; Round and Corbett, 2020). Our claim is narrower. Many NLP systems use typological databases after the database has already discretized a comparison. When one feature has several possible values, those values should usually be modeled as competing alternatives, not as unrelated binary labels.

We keep correspondence and association separate. A *correspondence* is a linguistically justified mapping between database values that encode the same, nearly the same, or explicitly decomposable construct. An *association* is any distributional resemblance, whether caused by equivalence, genealogy, area, diachrony, or sampling. Retrieval methods rank associations; the benchmark asks whether the top-ranked items recover curated correspondences. Observed profiles are the most reliable first-stage retrieval signal. Categorical embeddings are useful after lookup, when the question is how the values of one feature relate to one another.

The paper makes three contributions. First, it releases a typed WALS–GRAMBANK value crosswalk and evaluates cross-database retrieval against observed-profile and genealogy-only baselines. Second, it compares independent binary supervision with a categorical factorization that treats values as alternatives within a feature. Third, it shows that categorical value embeddings recover linguistically meaningful feature-internal structure; controlled prediction and downstream checks test the same claim from two additional angles.

2 Background

From labels to contrast sets. Independent binary expansion is usually chosen for engineering convenience, not because it reflects the structure of typological features. A categorical representation asks one question per feature and lets the values compete. The prediction effect is immediate: the model cannot assign full probability to mutually incompatible values. For interpretation, value embeddings can encode shared subproperties such as verb-finality, article presence, or the number of gender distinctions.

Database design and non-isomorphism. The WALS–GRAMBANK contrast is partly a difference in database format, not merely a difference in coverage. WALS is atlas-like. Many chapters assign each language one dominant or diagnostic value for a broad typological domain (Haspelmath et al., 2005; Dryer and Haspelmath, 2013). GRAMBANK is questionnaire-like. Most parameters ask whether a specific grammatical element, strategy, or construction is present or available. This design favors clear coding, computational usability, and fewer logical dependencies among features (Skirgård et al., 2023; Haynie et al., 2023). Thus WALS 81A classifies dominant subject-object-verb order, whereas GRAMBANK represents comparable clause-order evidence through more specific parameters such as verb-initial, verb-medial, and verb-final order. Converting one format into the other can lose information: WALS may hide secondary strategies, while many GRAMBANK parameters do not encode an explicit n-way contrast. Appendix D gives the longer design rationale.

This is why the crosswalk is typed rather than flat. Cross-resource feature comparison is already recognized as a methodological problem for typology-aware NLP and grammar-engineering resources (de Almeida et al., 2019). The mismatches vary by domain: articles often develop from demonstratives or the numeral 'one' (Diessel, 1999; Heine and Kuteva, 2002); classifiers differ across numeral, noun, possessive, and verbal systems (Aikhenvald, 2000); agreement depends on controller, target, feature, and domain (Corbett, 2006, 2022); and predicative possession includes existential, locational, comitative, transitive, and topic-like strategies (Stassen, 2009; Chappell and Creissels, 2019). A high profile match can reflect true equivalence, a one-way implication, a decomposition relation, or a shared diachronic source. It can also reflect genealogy, area, or coverage bias.

Figure 1 compares the geometry induced by an independent-binary encoding (T-CF, right) and a categorical softmax encoding (Learned, middle) on three WALS features whose values have different internal structure. The main figure now shows the pairwise distance matrices directly; Appendix M gives the corresponding structural projections and the full combined view.

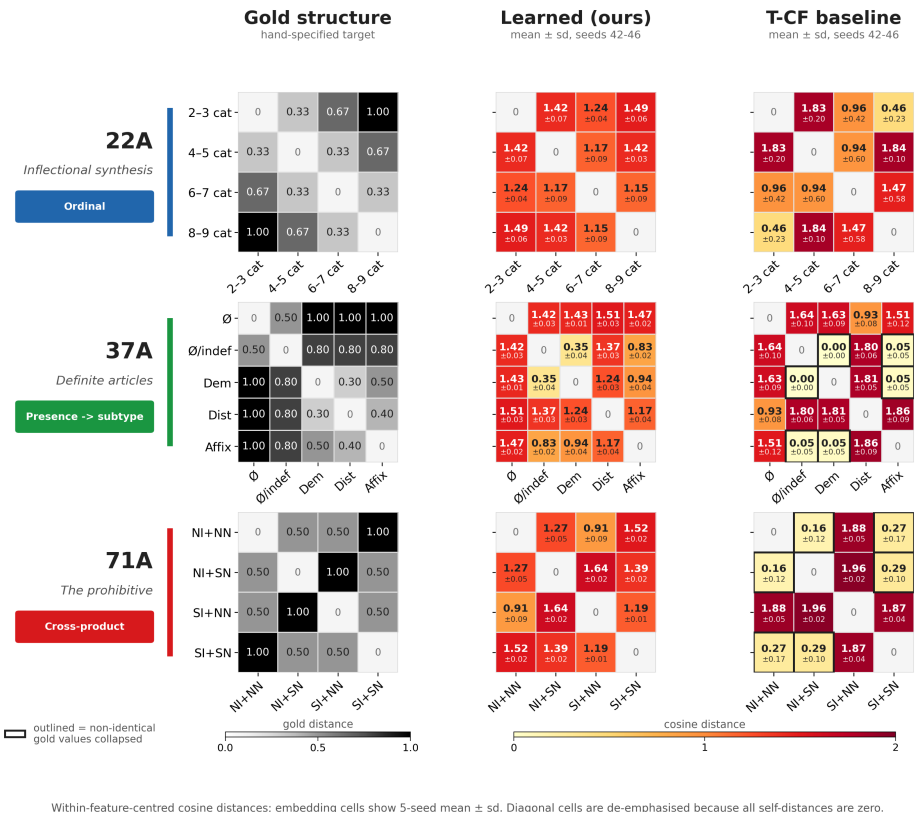

Figure 1: Central within-feature structure test, shown as distance matrices. For each displayed WALS feature, the left matrix is the annotated gold structure, the middle matrix is the five-run mean distance matrix for Learned, and the right matrix is the corresponding T-CF baseline. Rows represent an ordinal scale (22A), a presence-to-subtype system (37A), and a cross-product contrast (71A). Distances are computed after within-feature centering; values above 1.0 reflect negative cosine similarity and should be read ordinally. Categorical supervision more closely matches the annotated structure, while independent-binary T-CF can reverse ordinal direction, collapse non-identical article values, or weaken one axis of a cross-product contrast. Appendix M gives the projection view.

Computational typology and feature completion. NLP systems have used typology in several ways: to describe languages, guide transfer, probe representations, and fill missing database cells (Bender, 2009, 2016; Littell et al., 2017; Malaviya et al., 2017; Ponti et al., 2019; Bjerva et al., 2019c; Bjerva, 2024). These questions also predate recent neural embedding work. Earlier work compared typological and genealogical groupings (Georgi et al., 2010), learned continuous typological spaces for phylogenetic inference (Murawaki, 2015), identified informative typological features (Takamura et al., 2016), and treated missing cells as a matrix-completion problem (Buis and Hulden, 2019). Diachronic models of vertical stability and horizontal diffusibility make the same warning explicit: genealogy and contact are both signal and confound (Yamauchi and Murawaki, 2016; Murawaki and Yamauchi, 2018).

Language embeddings and typological probing. Prior NLP work has often represented languages with vectors. Some vectors are built from typological databases, as in URIEL/lang2vec and URIEL+ (Littell et al., 2017; Khan et al., 2025); related work completes missing typological features (Amirzadeh et al., 2025). Others are learned from multilingual tasks and then probed for typological structure (Östling and Tiedemann, 2017; Bjerva and Augenstein, 2018; Bjerva et al., 2019b; Choenni and Shutova, 2022; Östling and Kurfalı, 2023). These representations can capture genealogy, geography, and structural similarity, but those signals are often confounded. We therefore keep three objects separate: language embeddings, observed value profiles over languages, and embeddings of the feature-values themselves.

Universals and dependency transfer. Prior work has also tried to discover or score typolog-

ical implications from databases (Daumé III and Campbell, 2007; Bjerva et al., 2019a; Dönicke et al., 2020). Word order also matters for downstream syntax: typology has been used for selective sharing, target-language adaptation, multilingual parsing, and word-order transfer (Naseem et al., 2012; Täckström et al., 2013; Ammar et al., 2016; Gulordava and Merlo, 2016; Fisch et al., 2019; Liu et al., 2020; Üstün et al., 2022; Li et al., 2025). Our study asks an earlier question. Before a system transfers or completes typological values, which values should be linked, and what structure do those values have?

3 Task, Data, and Models

Typed cross-database alignment. For each WALS value, the retrieval task ranks candidate GRAMBANK atoms or formulas. A candidate is correct only if it appears in our high-confidence Tier-A crosswalk. Tier-B projection/directional rows and Tier-C partial overlaps are used only for error analysis, not as positive labels in the main retrieval evaluation. The main table reports R@1, R@5, R@10, and MRR; additional retrieval metrics are reported in the appendix.

Data. We use WALS CLDF v2020.4 (Dryer and Haspelmath, 2013; Forkel et al., 2018) and GRAMBANK CLDF v1.0 (Skirgård et al., 2023), intersected with BibleNLP/eBible (Akerman et al., 2023) only as a language filter for comparability with prior typological embedding work. No textual features enter the typological embedding models. After filtering for branch size, language coverage, and feature-value frequency, the training matrices contain 1,641 WALS languages (640 feature-values, 81 Glottolog branches) and 2,360 GRAMBANK languages (392 feature-values, 52 branches). Thus numbers such as the 186 binary GRAMBANK features discussed below refer only to the filtered matrix used in our experiments, not to the full 195-feature Grambank resource. The 1,015 shared Glottocodes (Hammarström et al., 2024) define the alignment space.

Crosswalk. The supplementary crosswalk contains 99 accepted relations across 32 WALS features, each mapping a WALS value to a GRAMBANK atom or formula. Rows are typed by relation, confidence tier, caveat, evidence type, semantic rationale, and observed-profile diagnostics. The corrected Tier-A retrieval set contains 74 pairs. To compare methods fairly, the main table uses only the 70 Tier-A pairs that all methods can score. The remaining Tier-A rows are reported in the supplementary materials. Appendix J gives the relation counts and correction log.

Observed profiles and confounds. For each value v, the observed profile $o_v \in \{0, 1, \mathrm{NaN}\}^{1015}$ records whether each shared language has the value, has a different value for the parent feature, or lacks attestation. We compare profiles using pairwise-deleted cosine, Matthews/ϕ (Matthews, 1975; Chicco and Jurman, 2020), Jaccard, positive pointwise mutual information (PPMI) (Church and Hanks, 1990), and smoothed log-odds over the 2×2 co-attestation table (Anscombe, 1956; Monroe et al., 2008). We add two genealogy-only baselines: family-prevalence cosine collapses each profile to prevalence over Glottolog families, and URIEL+ phylogenetic centroid cosine represents each value by the mean phylogenetic vector of languages that carry it.

Binary and categorical matrix factorization. The T-CF baseline adapts the independent-binary objective associated with Bjerva et al. (2019c): multi-valued features are expanded into independent one-hot columns, each modelled as $p(Y_{ij} = 1) = \sigma(\ell_i^\top f_j)$. The categorical model instead represents feature j with K_j values as

$$p(Y_{ij} = k \mid \ell_i) = \mathrm{softmax}_k(\ell_i^\top V_k^{(j)}). \quad (1)$$

The comparison is not mainly about model size: the two models have similar parameter counts, differing by less than 10%. Appendix E reports the parameter counts and the behavior of retained binary features.

Predicted profiles, rotation, and universals. For each value v, the predicted profile $\pi_v \in [0, 1]^{1015}$ gives the model's predicted probability for each shared language. Predicted profiles are compared by cosine. As a negative control, we also rotate value embeddings directly using a language-space Procrustes map. This negative control performs poorly, as expected if language and value embeddings are not directly aligned (Appendix Q). Appendix P describes the secondary universals probe derived from the GRAMBANK operationalizations of Verkerk et al. (2026).

4 Experiments and Results

We report retrieval first, then representation geometry, followed by two supporting checks. First, we ask which score best retrieves curated WALS–GRAMBANK correspondences. We then ask whether learned vectors recover structure inside a feature's contrast set. Controlled prediction and downstream linearization provide supporting checks; the universals and conditioning probes are reported in the appendix.

4.1 Exact cross-database lookup: observed profiles win

Table 1 reports retrieval on the 70-pair Tier-A subset scored by all methods. Using the same denominator prevents coverage differences from making some methods appear easier to evaluate than others. Observed-profile baselines are the strongest first-stage matchers. Observed ϕ has the best broad recall (R@10=.814), while Jaccard has the best first-rank precision and MRR (R@1=.614, MRR=.676). Cosine and smoothed log-odds both reach R@10=.771. PPMI is weaker, consistent with its sensitivity to rare perfect co-occurrences.

The same table also shows that genealogy is a strong confound. Family-prevalence cosine reaches R@10=.757, although it trails observed Jaccard and ϕ in MRR and first-rank precision. URIEL+ phylogenetic centroids are lower (R@10=.514). Aggregate retrieval scores are not enough. Each proposed correspondence still needs row-level inspection.

Neural predicted-profile retrieval is weaker than observed retrieval, but still informative. Learned predicted profiles exceed T-CF by a large margin (R@10=.371 vs. .100; MRR=.185 vs. .045), while direct value-space rotation remains at zero recall through R@10. Thus, categorical embeddings should not replace observed profiles for exact lookup. Their advantage is representational: they preserve more structure than independent binary expansions.

4.2 Representation geometry: categorical embeddings win

Retrieval scores alone do not test the main representational claim. If values are modeled as alternatives in one feature, related values should appear closer in the learned space. Figure 1 gives the main visual test. We annotate 70 WALS features with substructure schemas and evaluate ordinal rank, presence-subtype, Boolean-lattice, cross-product, and strategy-family diagnostics; the full protocol and table are in Appendix L. All diagnostics use within-feature-centred embeddings, which removes feature-level offsets and focuses the analysis on relations among values.

Method	R@1	R@5	R@10	MRR	n
Obs ϕ	.557	.771	.814	.653	70
Obs Jaccard	.614	.757	.771	.676	70
Obs cosine	.571	.729	.771	.642	70
Log-odds	.443	.714	.771	.555	70
PPMI	.157	.486	.629	.299	70
Family-prev. cos.	.500	.671	.757	.586	70
URIEL+ centroid	.257	.486	.514	.351	70
Learned predicted	.100	.300	.371	.185	70
T-CF predicted	.014	.071	.100	.045	70
Procrustes A	.000	.000	.000	.007	70

Table 1: Cross-database retrieval on the 70-pair Tier-A subset scored by all methods. All ten methods use the same denominator. Observed, PPMI/log-odds, family-prevalence, and URIEL+ rows score CLDF profiles over the 1,015 shared Glottocodes; neural rows average five independently trained runs. Appendix K gives R@3, median rank, and details on the shared-denominator filter.

The three rows of Figure 1 show the main failure modes. For 22A, the independent-binary baseline reverses the ordinal synthesis scale: values at opposite ends of the gold scale are closer than adjacent values. For 37A, it collapses non-identical article values, while the categorical model keeps the absent/present boundary visible. For 71A, the categorical space preserves both axes of the prohibitive cross-product more clearly; T-CF recovers some imperative structure but weakens the normal/special negative-marking contrast. The full diagnostics extend this reading beyond the displayed examples: Learned recovers ordinal direction on 22A and 30A, separates 37A definite articles, and recovers cross-product axes for 71A, 84A, 88A, and 97A. Lattice diagnostics are weaker, but still favor Learned over T-CF and the co-occurrence baseline.

Each diagnostic targets one kind of structure: order, absence versus subtype, component reconstruction, or recoverable cross-product axes. The full 70-feature output is included in the supplementary package.

Across features, Learned recovers familiar typological neighbourhoods. In WALS Learned, 81A=SOV is closest to OV+postpositions, OV

Matched experiment	Binary	Categorical	Δ
SIGTYP UFAL neural retrofit	0.567	0.636	+0.069
WALS F1 ($\rho = .20$)	0.580	0.622	+0.042
Grambank F1 ($\rho = .20$)	0.735	0.752	+0.017

Table 2: Predictive checks under matched conditions. The SIGTYP row is a controlled UFAL-style neural-component ablation: the embedding/scoring setup is held fixed while the supervision objective and output normalization are changed. The leave-one-branch-out rows are internal sanity checks under public data.

Representation / setting	τ	Pair	Arc
Seen: gold-profile oracle	0.402	–	–
Seen: WALS C hard	0.278	–	–
Seen: WALS B expected	0.272	–	–
Seen: Grambank A lang. emb.	0.270	–	–
Cold: Grambank B masked50	0.294	0.647	0.697
Cold: WALS B masked30	0.291	0.645	0.698
Cold: WALS B expected	0.287	0.643	0.698
Cold: Grambank B expected	0.276	0.638	0.695
Cold: WALS C hard	0.185	0.593	0.517
Cold: Grambank C hard	0.160	0.580	0.543

Table 3: Typology-conditioned dependency linearization. The cold block evaluates Belarusian after holding out WALS 81A "No dominant order". Expected categorical embeddings outperform hard profiles in this targeted cold-value setting.

order, postpositions, no SNegOV, and genitive-noun order, matching Greenbergian/Dryer head-final structure (Greenberg, 1963; Dryer, 1992). Ergative-absolutive alignment is close across 98A and 99A and near oblique-patient constructions. In the learned space, complex tone appears near isolating and zero-marking features often associated with Southeast Asian languages. This neighbourhood may reflect areal structure, genealogy, coverage, or all three. T-CF preserves some neighbours but compresses top cosines, reducing discriminability.

4.3 Controlled predictive checks

We first use leave-one-branch-out feature prediction to verify model behavior under the same public-data setting. Learned exceeds T-CF only after moderate supervision, and the frequency baseline is strong for GRAMBANK because many binary features have dominant majority classes. Because public WALS sources yield a much higher frequency baseline than the original T-CF paper, we use T-CF only as a matched internal control and document the audit in Appendix F.

A more controlled external comparison is a retrofit of the neural component of the UFAL system that won the SIGTYP 2020 Shared Task (Bjerva et al., 2020; Vastl et al., 2020). UFAL's neural component uses language and feature-value embeddings, dropout, a dot product, and a sigmoid output; our controlled reimplementation keeps the embedding/scoring setup fixed, then replaces independent binary sigmoid supervision and output normalization with feature-wise categorical softmax. Macro accuracy rises from 0.567 to 0.636 (+0.069). We use this only as an objective/output-normalization ablation of a UFAL-style neural component, not as a full reproduction of UFAL's submitted combined system or as a new state-of-the-art comparison.

4.4 Downstream proof of concept

To probe whether value geometry helps in a controlled sentence-level setting, we introduce typology-conditioned dependency linearization. Given an unordered UD tree, dependency labels, UPOS tags, and a target-language representation, the model predicts whether each dependent precedes its head and converts arc preferences into a sentence-level order. The key comparison is between hard binary/argmax profiles and expected categorical value embeddings.

In seen-language self-linearization, hard and expected representations are close: WALS C reaches mean Kendall's $\tau = 0.278$, WALS B reaches 0.272, and GRAMBANK A reaches 0.270. Our main diagnostic condition is cold-value generalization. Holding out the WALS 81A value "No dominant order" and evaluating on Belarusian, expected embeddings reach $\tau = 0.276$–0.294 and arc accuracy 0.695–0.698, versus $\tau = 0.160$–0.185 and arc accuracy 0.517–0.543 for hard binary/argmax profiles. This supports a narrow claim: the categorical space helps in a setting where a downstream syntactic decision depends on similarity to an unseen alternative inside a structured contrast set.

4.5 Additional diagnostics

Appendices O and P report two secondary diagnostics consistent with the main interpretation. The universals probe shows higher structured implication scores for supported formulas than for unsupported formulas. The URIEL+/lang2vec conditioning results differ by resource: in this setup, external geo/phylogenetic information helps WALS most when in-branch evidence is sparse, while GRAMBANK changes only slightly after some in-

branch evidence is available.

5 Analysis and Discussion

The results separate two questions. Cross-database lookup is mostly a profile-matching problem. When a WALS value and a GRAMBANK value pick out nearly the same languages, observed ϕ, Jaccard, and related profile scores are preferable because their evidence can be inspected language by language. Learned vectors are weaker at this first-pass task.

The value-space question is different. Figure 1 shows what changes when values are trained as alternatives inside a feature: ordinal scales, absence–subtype systems, and cross-product values acquire interpretable geometry. The 70-feature diagnostic suite shows that these effects are not limited to the displayed cases. This is the main reason to use categorical value embeddings when the goal is to analyze relations among values, not just predict labels.

The predictive and downstream experiments support this interpretation, but only in controlled settings. The SIGTYP retrofit improves when independent binary supervision is replaced by feature-wise categorical softmax under matched conditions. The TCDL cold-value split improves when a held-out word-order value can be represented by its position among related alternatives rather than by a missing binary bit. This resembles a common low-resource case: some typological evidence is available, but the exact value needed by the model is missing or rare.

Genealogy and area remain central confounds. Database values reflect inheritance, contact, and areal ecology, so a useful representation should not erase family structure. At the same time, genealogy can make unrelated values look similar. The retrieval results therefore suggest a staged workflow: use transparent observed profiles and family-prevalence checks for crosswalk construction, then use categorical embeddings to inspect relations among the linked values.

Two secondary diagnostics point in the same direction. The universals probe derives 172 value-space formulas from the GRAMBANK operationalizations of 191 Universals Archive entries in Verkerk et al. (2026); supported formulas score higher than unsupported formulas, though coexistence universals and absence-conditioned antecedents remain difficult. Crosswalk errors are also informative. A low rank can mean that retrieval failed, but it can also reveal a mismatch in definition, polarity, constructional scope, or sample coverage. Article systems, agreement, and possession are typical cases where profile similarity must be read together with definitions and caveats.

Typological embeddings should be used as diagnostic tools, not as replacements for database values or expert analysis. They are most useful when the question is explicit. Exact lookup, imputation, substructure analysis, universals testing, and downstream transfer require different kinds of evidence.

Using the evidence. The safest use of these methods is staged. Start with observed profiles. A proposed WALS–GRAMBANK link should first be supported by a transparent co-attestation pattern over shared languages, preferably with row-level evidence that can be checked against definitions. Then compare that pattern with family-prevalence and phylogenetic baselines. If the family-only score is nearly as strong as the observed score, the link may still be correct, but the evidence is no longer purely structural. It may also be inherited, areal, or coverage-driven.

The typed crosswalk is meant to make this review possible. A near-exact row should be readable as a candidate correspondence. A directional row says something weaker: one value tends to imply or instantiate another, but the reverse is not guaranteed. A projection row records that one resource decomposes a constructional domain that the other keeps bundled. A partial row is not a positive label for exact retrieval; it is evidence that the two databases divide a typological domain differently. This distinction matters because a high profile score can look convincing even when the linguistic relation is not equivalence.

Figure 1 is the complementary test. It does not ask whether SOV retrieves the right GRAMBANK formula. It asks whether the model has learned that SOV, OVS, SVO, and VSO are not just separate labels but related answers to one word-order question. The distance matrices show the local pairwise relations directly; the projection panels in Appendix M provide a more compact visual summary of the same distances.

The same logic applies outside word order. In an ordinal feature, adjacent values should usually be closer than endpoints. In a presence-subtype feature, absence should be separated from

the present subtypes, but the present subtypes should not necessarily collapse into one point. In a cross-product feature, the model should recover both axes rather than only the most frequent one. These expectations are not universal laws; they are feature-specific hypotheses derived from the database definitions and typological literature. The 70-feature diagnostic suite makes those hypotheses explicit.

The result is therefore not that embeddings are a universal replacement for typological annotation. Observed profiles are better for first-pass lookup because they are simple and inspectable. Categorical embeddings are better after lookup, when the question becomes whether values have a sensible internal organization, whether sparse values can be interpreted through nearby alternatives, or whether a downstream model can use similarity among values that were not all observed in training.

The appendices provide the audit trail for this comparison: the correction log, retrieval variants, full substructure protocol, conditioning sweep, universals probe, and implementation details. These materials are included so that individual crosswalk decisions and diagnostic choices can be checked without changing the main comparison between profile evidence and categorical value geometry.

The staged view is also a safeguard against two common mistakes. The first is to treat a high-ranked pair as a database correction. The methods do not license that step. They produce candidates and diagnostics; changing a typological resource still requires checking the definitions, sources, and constructional scope. The second mistake is to treat a visually coherent embedding space as proof that the values have the same linguistic interpretation across resources. A neighborhood can be meaningful without being an equivalence class. Demonstratives, definite articles, and noun-phrase order can cluster for good diachronic or distributional reasons even when they should remain separate database values.

This is why the paper keeps profile evidence, definitions, and geometry separate rather than merging them into a single score. A row is strongest when all three agree: the definitions match, the observed profile is high after genealogical checks, and the learned neighborhood is typologically sensible. A row is still useful when they disagree, but it should be read differently. A profile match with weak semantic support is a possible confound. A good semantic match with weak profile support may reflect sparse coverage or a difference in dominance coding. A good embedding neighborhood with weak exact retrieval may indicate a broader typological relation rather than a cross-database correspondence.

For that reason, the release is intended as an auditable starting point. The crosswalk gives candidate links and caveats. The retrieval files show which links are easy or difficult under observed profiles, count-based scores, genealogy baselines, and predicted profiles. The substructure files show which features behave as ordered scales, subtype systems, lattices, or cross-products. This makes the resource revisable without changing the paper's main comparison between independent binary and categorical supervision.

6 Conclusion

Multi-valued typological database features are often better modeled as choices among alternatives than as independent yes/no labels. For exact WALS–GRAMBANK lookup, observed ϕ, Jaccard, cosine, and log-odds give the most reliable first-stage signals, and the family-prevalence baseline shows why even strong profile matches require genealogical checks. For representation, categorical embeddings capture structure that observed profiles do not directly model. They recover relations among values of the same feature and help when a downstream model must generalize to a held-out typological value. Continuous vectors should not replace database values or typological analysis. Their value is to make relations among database values easier to inspect.

Limitations

Crosswalk and interpretation. The crosswalk is documented and typed, but it is not a community standard. Difficult domains such as causatives, possession, numeral systems, evidentiality, negation, and ergativity require independent typological review. The methods rank associations and expose structure; they do not determine typological validity.

Evaluation scope. The TCDL experiment is a targeted proof of concept, with the strongest cold-value result coming from one held-out WALS value and one language. The 70-feature substructure annotation reflects one author's reading of the typological literature, checked against WALS chapter descriptions; medium-confidence cells should be treated as hypotheses.

Baselines and conditioning. T-CF is used as an internal matched baseline because our audit did not recover the original low-frequency-baseline setting from public WALS sources. The SIGTYP retrofit is likewise an objective ablation of a UFAL-style neural component, not a full reproduction of the submitted system. URIEL+/lang2vec conditioning is evaluated only for the Learned architecture, and the GRAMBANK split uses Grambank Family as a coarse genus equivalent.

Ethical Considerations

WALS and GRAMBANK are public resources, but their coverage reflects inequalities in language documentation. Co-distributional methods may therefore favor well-documented families and macroareas. The proposed mappings should be used as diagnostic aids for cross-database review, not as replacements for descriptive expertise, community knowledge, or typological judgment.

All external resources are used under documented research licenses. The WALS CLDF snapshot and GRAMBANK CLDF v1.0 are distributed under CC BY 4.0 (CLDF Datasets, 2020; Grambank Consortium, 2023). Universal Dependencies treebanks carry per-treebank licenses, typically CC BY-SA 4.0 or compatible terms (Universal Dependencies Contributors, n.d.; Nivre et al., 2020). BibleNLP/eBible materials carry per-translation licenses and are used here only as a language-identity filter, with no textual features entering the models (Akerman et al., 2023; BibleNLP, 2023). URIEL+ is used under the license stated by the public repository (Lee Language Lab, 2025; Khan et al., 2025). The released crosswalk and code are distributed under CC BY 4.0 and MIT-compatible terms, respectively.

AI coding assistants were used during software development, and a large language model was used for editing assistance on draft text. The experimental design, implementation checks, reported results, analyses, and claims are the authors' own and were independently verified.

References

Vesa Akerman, David Baines, Damien Daspit, Ulf Hermjakob, Taeho Jang, Colin Leong, Michael Martin, Joel Mathew, Jonathan Robie, and Marcus Schwarting. 2023. The eBible corpus: Data and model benchmarks for Bible translation for low-resource languages. *arXiv:2304.09919*.

BibleNLP. 2023. eBible: Curated corpus of parallel data derived from versions of the Bible provided by eBible.org. `https://github.com/BibleNLP/ebible`.

Waleed Ammar, George Mulcaire, Miguel Ballesteros, Chris Dyer, and Noah A. Smith. 2016. Many languages, one parser. *Transactions of the Association for Computational Linguistics*, 4:431–444.

F. J. Anscombe. 1956. On estimating binomial response relations. *Biometrika*, 43(3/4):461–464.

Alexandra Y. Aikhenvald. 2000. *Classifiers: A Typology of Noun Categorization Devices*. Oxford University Press.

Hamidreza Amirzadeh, Sadegh Jafari, Anika Harju, and Rob van der Goot. 2025. data2lang2vec: Data driven typological features completion. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 6520–6529.

Emi Baylor, Esther Ploeger, and Johannes Bjerva. 2023. The past, present, and future of typological databases in NLP. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1163–1169.

Emi Baylor, Esther Ploeger, and Johannes Bjerva. 2024. Multilingual gradient word-order typology from Universal Dependencies. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics*, pages 42–49.

Emily M. Bender. 2009. Linguistically naïve $\neq$ language independent: Why NLP needs linguistic typology. In *Proceedings of the EACL 2009 Workshop on the Interaction between Linguistics and Computational Linguistics*.

Emily M. Bender. 2016. Linguistic typology in natural language processing. *Linguistic Typology*, 20(3):645–660.

Emily M. Bender and Batya Friedman. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604.

Balthasar Bickel. 2007. Typology in the 21st century: Major current developments. *Linguistic Typology*, 11(1):239–251.

Balthasar Bickel and Johanna Nichols. 2002. Autotypologizing databases and their use in fieldwork. In *Proceedings of the International LREC Workshop on Resources and Tools in Field Linguistics*, pages 33–40.

Johannes Bjerva. 2024. The role of typological feature prediction in NLP and linguistics. *Computational Linguistics*, 50(2):781–794.

Johannes Bjerva and Isabelle Augenstein. 2018. From phonology to syntax: Unsupervised linguistic typology at different levels with language embeddings. In *Proceedings of NAACL-HLT 2018*, pages 907–916.

Johannes Bjerva, Yova Kementchedjhieva, Ryan Cotterell, and Isabelle Augenstein. 2019a. Uncovering probabilistic implications in typological knowledge bases. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3924–3930.

Johannes Bjerva, Robert Östling, Maria Han Veiga, Jörg Tiedemann, and Isabelle Augenstein. 2019b. What do language representations really represent? *Computational Linguistics*, 45(2):381–389.

Johannes Bjerva, Yova Kementchedjhieva, Ryan Cotterell, and Isabelle Augenstein. 2019. A probabilistic generative model of linguistic typology. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1529–1540.

Johannes Bjerva, Elizabeth Salesky, Sabrina J. Mielke, Aditi Chaudhary, Giuseppe G. A. Celano, Edoardo Maria Ponti, Ekaterina Vylomova, Ryan Cotterell, and Isabelle Augenstein. 2020. SIGTYP 2020 shared task: Prediction of typological features. In *Proceedings of the Second Workshop on Computational Research in Linguistic Typology*, pages 1–11.

Annebeth Buis and Mans Hulden. 2019. Typological feature prediction with matrix completion. In *Proceedings of TyP-NLP: The First Workshop on Typology for Polyglot NLP*, pages 13–15.

Hilary Chappell and Denis Creissels. 2019. Topicality and the typology of predicative possession. *Linguistics*, 57(3):467–532.

Kenneth Ward Church and Patrick Hanks. 1990. Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16(1):22–29.

Rochelle Choenni and Ekaterina Shutova. 2022. Investigating language relationships in multilingual sentence encoders through the lens of linguistic typology. *Computational Linguistics*, 48(3):635–672.

Davide Chicco and Giuseppe Jurman. 2020. The advantages of the Matthews correlation coefficient over F1 score and accuracy in binary classification evaluation. *BMC Genomics*, 21:6.

Greville G. Corbett. 2006. *Agreement*. Cambridge University Press.

Greville G. Corbett. 2022. The agreement hierarchy revisited. *Languages*, 7(4):283.

William Croft. 2003. *Typology and Universals*, second edition. Cambridge University Press.

Hal Daumé III and Lyle Campbell. 2007. A Bayesian model for discovering typological implications. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 65–72.

Tifa de Almeida, Youyun Zhang, Kristen Howell, and Emily M. Bender. 2019. Feature comparison across typological resources. In *Proceedings of TyP-NLP: The First Workshop on Typology for Polyglot NLP*, pages 16–17.

Holger Diessel. 1999. *Demonstratives: Form, Function, and Grammaticalization*. John Benjamins.

Tillmann Dönicke, Xiang Yu, and Jonas Kuhn. 2020. Real-valued logics for typological universals: Framework and application. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3990–4003.

Matthew S. Dryer. 1992. The Greenbergian word order correlations. *Language*, 68(1):81–138.

Matthew S. Dryer and Martin Haspelmath, editors. 2013. *WALS Online*. Max Planck Institute for Evolutionary Anthropology.

CLDF Datasets. 2020. The World Atlas of Language Structures, CLDF dataset v2020.4. `https://github.com/cldf-datasets/wals`. License: CC BY 4.0.

Adam Fisch, Jiang Guo, and Regina Barzilay. 2019. Working hard or hardly working: Challenges of integrating typology into neural dependency parsers. In *Proceedings of EMNLP-IJCNLP 2019*, pages 5714–5720.

Ryan Georgi, Fei Xia, and William Lewis. 2010. Comparing language similarity across genetic and typologically-based groupings. In *Proceedings of COLING 2010*, pages 385–393.

Robert Forkel, Johann-Mattis List, Simon J. Greenhill, Christoph Rzymski, Sebastian Bank, Michael Cysouw, Harald Hammarström, Martin Haspelmath, Gereon A. Kaiping, and Russell D. Gray. 2018. Cross-linguistic data formats, advancing data sharing and re-use in comparative linguistics. *Scientific Data*, 5:180205.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. 2021. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92.

Joseph H. Greenberg. 1963. Some universals of grammar with particular reference to the order of meaningful elements. In Joseph H. Greenberg, editor, *Universals of Language*, pages 73–113. MIT Press.

Kristina Gulordava and Paola Merlo. 2016. Multi-lingual dependency parsing evaluation: A large-scale analysis of word order properties using artificial data. *Transactions of the Association for Computational Linguistics*, 4:343–356.

Harald Hammarström, Robert Forkel, Martin Haspelmath, and Sebastian Bank. 2024. *Glottolog 5.0*. Max Planck Institute for Evolutionary Anthropology.

Martin Haspelmath. 2010. Comparative concepts and descriptive categories in cross-linguistic studies. *Language*, 86(3):663–687.

Martin Haspelmath, Matthew S. Dryer, David Gil, and Bernard Comrie, editors. 2005. *The World Atlas of Language Structures*. Oxford University Press.

Hannah J. Haynie, Damián E. Blasi, Hedvig Skirgård, Simon J. Greenhill, Quentin D. Atkinson, and Russell D. Gray. 2023. Grambank's typological advances support computational research on diverse languages. In *Proceedings of the 5th Workshop on Research in Computational Linguistic Typology and Multilingual NLP*, pages 147–149.

Bernd Heine and Tania Kuteva. 2002. *World Lexicon of Grammaticalization*. Cambridge University Press.

Aditya Khan, Mason Shipton, David Anugraha, Kaiyao Duan, Phuong H. Hoang, Eric Khiu, A. Seza Doğruöz, and En-Shiun Annie Lee. 2025. URIEL+: Enhancing linguistic inclusion and usability in a typological and multilingual knowledge base. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 6937–6952.

Lee Language Lab. 2025. URIELPlus: URIEL+ knowledge base for natural language processing. https://github.com/LeeLanguageLab/URIELPlus. License: CC BY-SA 4.0.

Natalia Levshina. 2019. Token-based typology and word order entropy: A study based on Universal Dependencies. *Linguistic Typology*, 23(3):533–572.

Natalia Levshina, Savithry Namboodiripad, Marc Allassonnière-Tang, Mathew Kramer, Luigi Talamo, Annemarie Verkerk, Sasha Wilmoth, Gabriela Garrido Rodriguez, Timothy Michael Gupton, Evan Kidd, Zoey Liu, Chiara Naccarato, Rachel Nordlinger, Anastasia Panova, and Natalia Stoynova. 2023. Why we need a gradient approach to word order. *Linguistics*, 61(4):825–883.

Zhuoran Li, Chunming Hu, Junfan Chen, Zhijun Chen, and Richong Zhang. 2025. Implicit word reordering with knowledge distillation for cross-lingual dependency parsing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(23):24530–24538.

Patrick Littell, David R. Mortensen, Ke Lin, Katherine Kairis, Carlisle Turner, and Lori Levin. 2017. URIEL and lang2vec: Representing languages as typological, geographical, and phylogenetic vectors. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, pages 8–14.

Lu Liu, Yi Zhou, Jianhan Xu, Xiaoqing Zheng, Kai-Wei Chang, and Xuanjing Huang. 2020. Cross-lingual dependency parsing by POS-guided word reordering. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2938–2948.

Chaitanya Malaviya, Graham Neubig, and Patrick Littell. 2017. Learning language representations for typology prediction. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2529–2535.

Brian W. Matthews. 1975. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochimica et Biophysica Acta*, 405(2):442–451.

Leland McInnes, John Healy, and James Melville. 2018. UMAP: Uniform manifold approximation and projection for dimension reduction. *arXiv:1802.03426*.

Burt L. Monroe, Michael P. Colaresi, and Kevin M. Quinn. 2008. Fightin' words: Lexical feature selection and evaluation for identifying the content of political conflict. *Political Analysis*, 16(4):372–403.

Yugo Murawaki. 2015. Continuous space representations of linguistic typology and their application to phylogenetic inference. In *Proceedings of NAACL-HLT 2015*, pages 324–334.

Yugo Murawaki and Kenji Yamauchi. 2018. A statistical model for the joint inference of vertical stability and horizontal diffusibility of typological features. *Journal of Language Evolution*, 3(1):13–25.

Tahira Naseem, Regina Barzilay, and Amir Globerson. 2012. Selective sharing for multilingual dependency parsing. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, pages 629–637.

Frederick J. Newmeyer. 2007. Linguistic typology requires crosslinguistic formal categories. *Linguistic Typology*, 11(1):133–157.

Johanna Nichols. 1992. *Linguistic Diversity in Space and Time*. University of Chicago Press.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. 2020. Universal Dependencies v2: An evergrowing multilingual treebank collection. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4034–4043.

Universal Dependencies Contributors. n.d. Licensing UD treebanks. https://universaldependencies.org/contributing/licensing.html.

Robert Östling and Jörg Tiedemann. 2017. Continuous multilinguality with language vectors. In *Proceedings of EACL 2017, Volume 2: Short Papers*, pages 644–649.

Robert Östling and Murathan Kurfalı. 2023. Language embeddings sometimes contain typological generalizations. *Computational Linguistics*, 49(4):1003–1051.

Frans Plank and Elena Filimonova. 2000. The Universals Archive: A brief introduction. *Sprachtypologie und Universalienforschung*, 53(1):109–123.

Edoardo Maria Ponti, Helen O'Horan, Yevgeni Berzak, Ivan Vulić, Roi Reichart, Thierry Poibeau, Ekaterina Shutova, and Anna Korhonen. 2019. Modeling language variation and universals: A survey on typological linguistics for natural language processing. *Computational Linguistics*, 45(3):559–601.

Erich R. Round and Greville G. Corbett. 2020. Comparability and measurement in typological science: The bright future for linguistics. *Linguistic Typology*, 24(3):489–525.

Erhard Rahm and Philip A. Bernstein. 2001. A survey of approaches to automatic schema matching. *VLDB Journal*, 10(4):334–350.

Pavel Shvaiko and Jérôme Euzenat. 2013. Ontology matching: State of the art and future challenges. *IEEE Transactions on Knowledge and Data Engineering*, 25(1):158–176.

Grambank Consortium. 2023. Grambank CLDF dataset v1.0. https://github.com/grambank/grambank. License: CC BY 4.0.

Hedvig Skirgård, Hannah J. Haynie, Damián E. Blasi, Harald Hammarström, Jeremy Collins, Jay J. Latarche, Jakob Lesage, Tobias Weber, Alena Witzlack-Makarevich, Sam Passmore, Angela Chira, Luke Maurits, Russell Dinnage, Michael Dunn, Ger Reesink, Ruth Singer, Claire Bowern, Patience Epps, Jane Hill, Outi Vesakoski, Martine Robbeets, Noor Karolin Abbas, Daniel Auer, Nancy A. Bakker, Giulia Barbos, Robert D. Borges, Swintha Danielsen, Luise Dorenbusch, Ella Dorn, John Elliott, Giada Falcone, Jana Fischer, Yustinus Ghanggo Ate, Hannah Gibson, Hans-Philipp Göbel, Jemima A. Goodall, Victoria Gruner, Andrew Harvey, Rebekah Hayes, Leonard Heer, Roberto E. Herrera Miranda, Nataliia Hübler, Biu Huntington-Rainey, Jessica K. Ivani, Marilen Johns, Erika Just, Eri Kashima, Carolina Kipf, Janina V. Klingenberg, Nikita König, Aikaterina Koti, Richard G. A. Kowalik, Olga Krasnoukhova, Nora L. M. Lindvall, Mandy Lorenzen, Hannah Lutzenberger, Tânia R. A. Martins, Celia Mata German, Suzanne van der Meer, Jaime Montoya Samamé, Michael Müller, Saliha Muradoglu, Kelsey Neely, Johanna Nickel, Miina Norvik, Cheryl Akinyi Oluoch, Jesse Peacock, India O. C. Pearey, Naomi Peck, Stephanie Petit, Sören Pieper, Mariana Poblete, Daniel Prestipino, Linda Raabe, Amna Raja, Janis Reimringer, Sydney C. Rey, Julia Rizaew, Eloisa Ruppert, Kim K. Salmon, Jill Sammet, Rhiannon Schembri, Lars Schlabbach, Frederick W. P. Schmidt, Amalia Skilton, Wikaliler Daniel Smith, Hilário de Sousa, Kristin Sverredal, Daniel Valle, Javier Vera, Judith Voß, Tim Witte, Henry Wu, Stephanie Yam, Jingting Ye, Maisie Yong, Tessa Yuditha, Roberto Zariquiey, Robert Forkel, Nicholas Evans, Stephen C. Levinson, Martin Haspelmath, Simon J. Greenhill, Quentin D. Atkinson, and Russell D. Gray. 2023. Grambank reveals the importance of genealogical constraints on linguistic diversity and highlights the impact of language loss. *Science Advances*, 9(16):eadg6175.

Leon Stassen. 2009. *Predicative Possession*. Oxford University Press.

Hiroya Takamura, Ryo Nagata, and Yoshifumi Kawasaki. 2016. Discriminative analysis of linguistic features for typological study. In *Proceedings of LREC 2016*, pages 69–76.

Oscar Täckström, Ryan McDonald, and Joakim Nivre. 2013. Target language adaptation of discriminative transfer parsers. In *Proceedings of NAACL-HLT 2013*, pages 1061–1071.

Universals Archive. n.d. Universals Archive. University of Konstanz. https://typo.uni-konstanz.de/archive/intro/.

Ahmet Üstün, Arianna Bisazza, Gosse Bouma, and Gertjan van Noord. 2022. UDapter: Typology-based language adapters for multilingual dependency parsing and sequence labeling. *Computational Linguistics*, 48(3):555–592.

Martin Vastl, Daniel Zeman, and Rudolf Rosa. 2020. Predicting typological features in WALS using language embeddings and conditional probabilities: ÚFAL submission to the SIGTYP 2020 Shared Task. In *Proceedings of the Second Workshop on Computational Research in Linguistic Typology*, pages 29–35.

Annemarie Verkerk, Olena Shcherbakova, Hannah J. Haynie, Hedvig Skirgård, Christoph Rzymski, Quentin D. Atkinson, Simon J. Greenhill, and Russell D. Gray. 2026. Enduring constraints on grammar revealed by Bayesian spatiophylogenetic analyses. *Nature Human Behaviour*, 10(1):126–136.

Kenji Yamauchi and Yugo Murawaki. 2016. Contrasting vertical and horizontal transmission of typological features. In *Proceedings of COLING 2016*, pages 836–846.

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605.

A Appendix Roadmap

The appendix gives the details needed to audit and reproduce the main results. Appendix B gives an extended analysis and discussion. Appendix C gives fuller limitations and ethics notes. Appendix D explains the WALS–GRAMBANK format difference in more detail. The implementation appendix describes filtering, optimization, parameter counts, and the behavior of binary features under the categorical objective. Later appendices report the predictive evaluation and replication audit for Bjerva et al. (2019c), the SIGTYP 2020 retrofit, typology-conditioned dependency linearization, crosswalk construction, retrieval and genealogy diagnostics, the within-feature substructure protocol behind Figure 1 and the corresponding distance matrices, cross-feature neighborhoods, family preservation, run stability, the URIEL+/lang2vec conditioning sweep, the universals probe, projection diagnostics, the tree-clustering baseline, and compute.

Each experiment targets a different question. Retrieval evaluates cross-database lookup. Substructure evaluates feature-internal structure. Universals test whether the value space reflects independently supported implications. Linearization asks whether value representations help in a sentence-level task. These experiments have different failure modes, so their scores should not be collapsed into a single measure of success.

B Extended Analysis and Discussion

This section gives the fuller analysis and discussion behind the compressed main-paper argument.

The retrieval table highlights a limitation. Exact cross-database lookup is largely a profile-matching problem. When a WALS value and a GRAMBANK value pick out nearly the same set of languages, observed ϕ or Jaccard is preferable because the evidence can be inspected language by

language. Learned vectors are weaker at this first-pass job.

That result does not settle the representation question. The stronger test is the one made visible in Figure 1: what happens inside a feature once its values are treated as alternatives rather than as independent columns. For 81A, the question is not only whether SOV retrieves the right Grambank formula, but whether SOV, SVO, VSO, and the rarer orders occupy positions that reflect shared subproperties. The 70-feature substructure analysis asks the same question more generally: whether ordinal values have ordinal distances, whether presence–subtype systems separate absence from subtypes, whether Boolean values reconstruct their components, and whether cross-product values recover their axes.

The three auxiliary experiments test the same claim from different angles. The SIGTYP retrofit shows a predictive gain when independent BCE and its output normalization are replaced by feature-wise cross-entropy under matched conditions. The substructure probe shows that the same objective yields more interpretable within-feature geometry. The TCDL cold-value split tests a downstream setting in which a model cannot simply read a known binary typological bit. No single diagnostic settles the issue. Together, they support a narrower claim: categorical feature-value embeddings help when a task requires interpolation among mutually exclusive typological alternatives.

The downstream result is narrow, but it is useful. In seen-language self-linearization, hard typological profiles remain competitive because the relevant values are already available. Expected categorical embeddings help most under cold-value generalization, where the held-out value has no active training bit but still has a position in the learned value space. This matches prior findings that typological information helps neural dependency parsing unevenly, often through coarse syntactic grouping or corpus-consistent statistics rather than through clean compositional use of each typological dimension (Fisch et al., 2019).

Genealogy and area remain both signal and confound. A typological representation with no recoverable family structure would be suspicious, since database values themselves reflect inheritance, contact, and areal ecology. The problem is that genealogy can also align unrelated values when they are frequent in the same families. The denominator-matched retrieval table makes this visible: family-prevalence cosine is a strong R@10 baseline, although observed ϕ and Jaccard remain better first-stage anchors by MRR and first-rank precision. The strongest cross-database evidence comes from rows where the semantic rationale is clear and the observed profile remains strong after comparison with family-prevalence and URIEL+ phylogenetic baselines. The conditioning sweep is compatible with this picture. External geo/phylogenetic information helps WALS most when in-branch evidence is extremely sparse, whereas GRAMBANK shows only small gains once some in-branch evidence is available. The projection and family-preservation diagnostics suggest why: GRAMBANK Learned spaces already carry stronger family and macroarea organization, while WALS leaves more room for external priors.

The universals probe adds a different check. Starting from the Verkerk et al. Grambank operationalizations of 191 Universals Archive entries, we derive 172 formulas that are testable in our learned value space. Supported formulas have much higher structured implication cosine than unsupported formulas, and T-CF yields no significant analogy residuals. This is not a claim that every co-occurrence is a universal. It shows that, in this probe, the categorical value space is closer to independently supported implication formulas than the binary control. Coexistence universals and absence-conditioned antecedents remain difficult because they do not always correspond to simple closeness among positive values.

Crosswalk errors are also informative. A low rank can indicate a failed retrieval, but it can also reveal a mismatch in ontology, granularity, polarity, constructional scope, or sample coverage. Article systems illustrate the point. For exact lookup, a definite-article value should retrieve GB020. In the learned space, the same value may sit near demonstratives or NP-order environments because those domains are related diachronically and constructionally. Agreement and possession show similar scope effects. The release therefore treats mappings as typed relations with caveats rather than as a flat list of positives.

These methods work best in stages. Cross-database integration should start with transparent observed-profile statistics and manual review of high-ranked candidates. Categorical embeddings are better after that anchoring step: they can smooth sparse values, propose typological neigh-

borhoods, expose feature-internal structure, and test whether similarity among alternatives helps downstream. Exact matching, imputation, substructure recovery, universals testing, and downstream transfer need separate evaluations; success in one task does not guarantee success in the others.

For NLP, the broader point is that typological information need not collapse into a static language vector. A feature value can have a definition, an observed profile, a cross-database formula, a set of nearest neighbors, a substructure diagnosis, and a downstream behavior. For typology, embeddings are useful only when the typological analysis supplies the questions they must answer. Comparative concepts then become objects of comparison in their own right.

C Extended Limitations and Ethics Notes

This section gives extended limitations and ethics notes.

Crosswalk scope. The crosswalk is documented and typed, but it is not a community standard. Each row includes a rationale, formula, caveat, and profile diagnostic, yet difficult domains still require independent typological review. The highest-risk areas are morphological causatives, possessive classification, numeral bases, evidentiality, negation morphology, and ergativity splits, where small changes in constructional scope can change the mapping.

Downstream linearization scope. The TCDL experiment is a targeted proof of concept rather than a full parsing evaluation. Its strongest result comes from one held-out WALS value, "No dominant order", and one language, Belarusian. Future work should add confidence intervals at the sentence level, relation-specific analyses, additional held-out values where UD coverage permits, stronger typology-free baselines, and full parsing experiments using relinearized treebanks or typology-conditioned distillation.

Substructure annotation. The 70-feature substructure annotation reflects one author's reading of the typological literature, checked against WALS chapter texts. Medium-confidence or provisional cells should be treated as hypotheses. Boolean-lattice diagnostics are underpowered when the number of values is small, and two-valued centroid-direction probes are not informative.

Replicability and baselines. The WALS T-CF result at $\rho = 20\%$ is far below the score reported by Bjerva et al. (2019c). Our audit did not recover the original low-frequency-baseline setting from public WALS sources. T-CF is therefore used as an internal matched baseline, not as a numerical replication. The SIGTYP retrofit should be interpreted similarly: it compares objectives under matched conditions, but it is not a full reproduction of UFAL's combined system.

Conditioning sweep. URIEL+/lang2vec conditioning is evaluated across three supervision fractions, five repeats, and geo/phylo/both variants, but only for the Learned architecture. The repeated branch-level splits are the relevant unit of comparison. The GRAMBANK split uses Grambank Family as a genus equivalent, which is coarser than the WALS genus split; the WALS–GRAMBANK contrast should therefore be read as a comparison between resource structures, not as a perfectly matched genealogical control. T-CF conditioning and more tightly matched ablations remain open.

Expert interpretation. The methods rank associations and reveal diagnostic structure. They do not determine typological validity. Candidate alignments should be checked against descriptive sources, expert judgment, and, where relevant, speaker or community knowledge before being used to revise database decisions.

Ethical considerations. WALS and GRAMBANK are public resources, but their coverage reflects long-standing inequalities in language documentation. Co-distributional methods work best for well-documented families and macroareas and may therefore reproduce those inequalities. The proposed mappings should be used as diagnostic aids: they can identify likely cross-database links, coverage gaps, and domains that need typological review, but they should not be used to overwrite descriptive expertise or community knowledge.

All external resources are used under their documented research licenses. The WALS CLDF snapshot used here is distributed under CC BY 4.0 (CLDF Datasets, 2020; Dryer and Haspelmath, 2013); GRAMBANK CLDF v1.0 is distributed under CC BY 4.0 (Grambank Consortium,

2023; Skirgård et al., 2023); and Universal Dependencies treebanks have per-treebank licenses, with CC BY-SA 4.0 as the default for new treebank repositories and individual exceptions documented by the data providers (Universal Dependencies Contributors, n.d.; Nivre et al., 2020). BibleNLP/eBible is used only as a language-identity filter: no Bible text or textual features enter the models, and the underlying translations retain the per-translation redistribution terms documented by the corpus and repository (Akerman et al., 2023; BibleNLP, 2023). URIEL+ is used under the CC BY-SA 4.0 license stated by the public repository (Lee Language Lab, 2025; Khan et al., 2025). The typed crosswalk and analysis code released with this paper are distributed as CC BY 4.0 and MIT-licensed artifacts, respectively, so that downstream users can inspect, revise, and reuse the resource while preserving attribution.

AI coding assistants were used during software development, and a large language model was used for editing assistance on draft text. The experimental design, implementation checks, reported results, analyses, and claims are the authors' own and were independently verified.

D Why WALS and GRAMBANK Differ in Format

The key consequence of the design contrast is that WALS values often behave as alternatives inside one comparative decision, while GRAMBANK parameters often behave as separate presence/absence traits. The fuller rationale clarifies why a typed crosswalk is needed.

WALS as atlas variables. WALS was built as a world atlas of structural features. A feature is a structural property used to distinguish languages by map values, and the chapter text defines how those values are assigned (Haspelmath et al., 2005; Dryer and Haspelmath, 2013). The atlas format encourages finite, visually distinguishable value sets. For example, 81A starts from the six logically possible orders of subject, object, and verb in transitive declarative clauses and adds a "no dominant order" value for languages where no single order is dominant. A language is therefore assigned a value for a chapter-level typological question. This is a categorical map-variable design: the values are alternatives for one comparison, even when some values encode absence, indeterminacy, mixture, or non-applicability.

Grambank as presence/absence traits. GRAMBANK was designed with a different operational logic. Its feature names are usually questions about whether a grammatical element, construction, or strategy is present or possible. The release paper and supplementary materials state that individual structural features were formulated mainly as binary yes/no values because this yields a simple and interpretable data structure, while a small set of multistate word-order features can be binarized for analysis (Skirgård et al., 2023). The NLP-facing description by Haynie et al. (2023) makes the computational rationale explicit: binary coding for most features and curation to avoid logical dependencies make the resource easier to use in computational applications. Presence coding also permits multiple strategies to be recorded for the same broad function, rather than forcing the database to choose only the most common or dominant strategy.

Why this matters for crosswalks and models. The two formats answer different questions. A WALS feature such as 81A asks which broad word-order type best characterizes a language; Grambank's corresponding clause-order parameters ask whether verb-initial, verb-medial, or verb-final orders are available under specified conditions. This makes partial annotation and multi-strategy coding easier in Grambank, but it also removes the within-feature competition structure that WALS makes explicit. Conversely, WALS preserves the n-way contrast but may hide secondary strategies. The crosswalk therefore cannot be a one-to-one table of names. Some rows are near-exact correspondences, others are decompositions or one-way implications, and still others are partial overlaps caused by different constructional scopes. This is why the main paper keeps a typed crosswalk and why the modelling comparison treats WALS-style multi-valued features as categorical contrast sets rather than as unordered bundles of independent binary targets.

E Implementation Details

Language set and filtering. WALS CLDF v2020.4 and Grambank CLDF v1.0 are intersected with BibleNLP/eBible only as a language filter for comparability with previous typological embedding work. No textual features, token sequences, or Bible-derived language representations enter the models. Language embeddings are learned

as ID lookup parameters from typological labels alone. Filtering follows the public-data setting in the main paper: branches with at least five languages; at least ten observed values per language; and at least ten languages per feature-value. After filtering, WALS has 1,641 languages, 640 feature-values, and 81 Glottolog branches; Grambank has 2,360 languages, 392 feature-values, and 52 branches. The cross-database alignment space consists of 1,015 shared Glottocode languages.

Optimization. All matrix-factorization models use embedding dimension 64, batch size 64, learning rate 10^{-3}, Adam with $\beta_1 = 0.9$ and $\beta_2 = 0.999$, early stopping on a 5% held-out validation split, patience 10, and a 200-epoch maximum. Sparse embedding updates led to undesirable optimization behavior with Adam's built-in weight decay: absent-batch embeddings received constant-rate sign-gradient descent and collapsed to near-zero norms. We therefore set Adam weight_decay=0 and add manual L2 regularization only over embeddings touched by the current batch. Without this change, the model can converge to a numerically stable solution in which the embedding norms have little substantive interpretation.

Parameter counts. The binary T-CF and categorical Learned models are close in capacity, so their comparison is about objective structure rather than parameter budget. Parameter counts are approximately: WALS T-CF 144k, WALS Learned 146k, Grambank T-CF 164k, and Grambank Learned 176k. The categorical model has additional rows for binary-feature opposites, but those rows do not create an uncontrolled capacity advantage.

Binary-feature geometry. For the 186 binary GRAMBANK features retained after our filtering and all 31 two-valued WALS features in these trained models, the categorical model typically converges to nearly antipodal value embeddings: $\cos(v_0, v_1) \approx -1.000$. This is consistent with the expected behaviour of a two-way softmax with symmetric regularization: a binary feature can be represented as one latent axis with opposite endpoints. Thus the same formulation covers binary features while avoiding one-vs-all expansion for multi-valued ones.

F Predictive Evaluation and Replication Audit

The predictive experiment is reported only to check that the models behave sensibly under the public-data setting. Table 4 gives the compact supervision grid used in the main paper; the full grid is included with the supplementary code.

ρ	Freq.	T-CF	Learned
WALS			
0%	.523	.328	.322
5%	.523	.551	.535
20%	.524	.580	.622
GRAMBANK			
0%	.722	.495	.498
5%	.722	.711	.717
20%	.723	.735	.752

Table 4: Mean F1 across branches and five repeats. The frequency baseline is strong at low supervision, especially for Grambank, where binary features often have dominant majority classes.

Why T-CF is an internal baseline, not a numerical replication. Bjerva et al. (2019) report a much higher WALS T-CF score than the public-data run here. We therefore treat T-CF as an architecturally matched control under the same public sources, preprocessing, splits, random repeats, and metrics as Learned, not as a faithful numerical replication of the NAACL 2019 experiment. The key discrepancy is the frequency baseline. Bjerva et al.'s reported frequency baseline is approximately .295, but all public WALS sources we tested yield approximately .52. The audit in Table 5 shows that the Bible intersection, WALS snapshot, threshold order, and training length do not recover the original low-frequency regime.

G SIGTYP 2020 Categorical Retrofit

The retrofit tests whether the categorical objective helps in a previously successful typological-prediction setup. We re-implement the neural component of the UFAL system, which won the SIGTYP 2020 Shared Task on prediction of typological features (Bjerva et al., 2020; Vastl et al., 2020). UFAL's neural component represents languages and feature-values with embedding tables, applies dropout, computes a dot product between the two embeddings, and feeds the result to a sigmoid layer before predicting whether a WALS feature-value belongs to a language (Vastl et al., 2020). This corresponds to the independent value-indicator setting studied here.

Audit item	Result
Bjerva et al. reference	Frequency baseline $\approx .295$; T-CF at $\rho = .20 \approx .88$.
Public WALS CLDF	Frequency baseline $\approx .519$ over 81 branches.
Bible-filtered public WALS	Frequency baseline $\approx .526$; the Bible intersection is not explanatory.
Public WALS snapshots	CLDF v2013/v2014/v2020.4, Zenodo v2013, and wals3 v2014.1 all give majority baselines around .52.
Threshold order	Applying the ≥ 10 threshold inside the Bible subset gives $\approx .523$.
Training length	Slavic $\rho = .20$: epoch 10 $\approx .717$, epoch 30 $\approx .751$, epoch 200 $\approx .724$.

Table 5: Replication-gap audit for the Bjerva et al. (2019) T-CF baseline. Public sources do not recover the original low-frequency-baseline setting; all comparisons in this paper are therefore internal matched comparisons.

Our retrofit is intended as a controlled contrast, not a full reproduction of UFAL's submitted ensemble: independent Bernoulli BCE and the corresponding output normalization are replaced with a feature-wise softmax and cross-entropy over the K_j candidate values of feature j. Embedding tables, dot-product scoring, optimizer, hyperparameters, language/feature-value vocabularies, and SIGTYP 2020 train/dev/test splits are fixed in our implementation. This isolates the supervision objective from architecture and capacity within the controlled reimplementation.

Objective	Macro acc.
Binary sigmoid (BCE)	.567
Categorical softmax (CE)	.636
Δ	+.069

Table 6: Macro-averaged accuracy on the SIGTYP 2020 test data for the matched objective comparison. The absolute binary baseline is below UFAL's reported Neural and Combined scores because this is a re-implementation of the neural component only, without their conditional-probability ensemble. The relevant quantity is the within-experiment objective gap.

The result complements the substructure analysis: in this matched run, changing the objective and output normalization improves external typological prediction and also produces value geometries that better reflect ordinal, tree, lattice, and cross-product structure. The categorical objective is therefore not only an interpretability device; under this controlled objective ablation, it also improves a prediction setting where multi-valued features are evaluated as multi-valued features.

H Typology-Conditioned Dependency Linearization

The typology-conditioned dependency linearization (TCDL) experiment asks whether feature-value geometry helps outside typological feature prediction. Given an unordered Universal Dependencies tree, UPOS tags, dependency labels, and a target-language typological representation T_ℓ, the model predicts pairwise head-dependent direction:

$$p(d < h \mid r, u_h, u_d, T_\ell) = \sigma\big(g_\theta([e_r; e_{u_h}; e_{u_d}; T_\ell])\big). \quad (2)$$

where r is the dependency relation and u_h, u_d are head and dependent UPOS tags. Pairwise preferences are converted into a sentence-level token ranking and evaluated with Kendall's τ, pair accuracy, and arc accuracy.

We compare three representation families. A is the learned language embedding from typological factorization. B is an expected categorical feature-value representation under the model's predicted value probabilities. C is a hard binary/argmax typological profile. The key contrast is B vs. C under cold-value conditions: hard profiles can retain many observed typological bits, but they cannot encode graded similarity between a held-out value and observed alternatives of the same feature.

The experiment targets a narrow setting. It does not show that learned typological embeddings dominate every multilingual transfer setting. It suggests that they can help in the kind of setting predicted by the theory: when a downstream syntactic decision must interpolate over an unseen or rare alternative inside a structured typological contrast set.

I Typed Crosswalk Construction

The crosswalk is a typed relation table, not a flat list of gold pairs. Cross-database alignment is closer to schema or ontology matching than to ordinary missing-value prediction (Rahm and Bernstein, 2001; Shvaiko and Euzenat, 2013). A row is accepted only when a WALS value and a Grambank atom or formula have a documented relationship; automatic statistics rank and audit candidates, but they do not assign labels.

Condition / representation	Kendall τ	Pair acc.	Arc acc.
Seen-language self-linearization, 41 languages			
Gold-profile oracle	.402	–	–
WALS C (binary/argmax)	.278	–	–
WALS B (expected)	.272	–	–
Grambank A (language embedding)	.270	–	–
WALS B masked30	.269	–	–
Grambank B (expected)	.238	–	–
Random	-.002	–	–
Cold value: Belarusian, WALS 81A "No dominant order" held out			
Grambank B masked50	.294	.647	.697
WALS B masked30	.291	.645	.698
WALS B (expected)	.287	.643	.698
Grambank B (expected)	.276	.638	.695
WALS C (binary/argmax)	.185	.593	.517
Grambank C (binary/argmax)	.160	.580	.543

Table 7: Typology-conditioned dependency linearization. Seen-language self-linearization shows that hard profiles remain competitive when the relevant word-order value is observed. The cold-value Belarusian split is diagnostic: expected categorical embeddings preserve similarity among alternatives of WALS 81A and substantially outperform hard binary/argmax profiles.

Each row records: WALS feature ID, WALS value label, Grambank formula, relation type, confidence tier, scope caveat, evidence type, semantic rationale, and observed-profile diagnostics. The supplementary crosswalk has 99 accepted relations across 32 WALS features: 72 NEAR-EXACT, 17 DIRECTIONAL, 6 PROJECTION, and 4 PARTIAL rows. Tier A contains 76 high-confidence rows. The main comparison uses the 70-pair subset scored by all ten methods; the remaining corrected Tier-A rows are retained for per-method diagnostics. Tier B contains 19 caveated rows; Tier C contains 4 analysis-only partial overlaps.

Relation types. NEAR-EXACT rows encode the same condition with at most a documented scope gap. DIRECTIONAL rows encode one-way implications, typically because WALS codes dominance and Grambank codes presence. PROJECTION rows collapse several WALS values into the same Grambank formula or vice versa. PARTIAL rows overlap in domain but differ in scope, word/affix treatment, granularity, or constructional coverage. Only Tier-A rows are positives for the main R@k/MRR retrieval evaluation; Tier-B and Tier-C rows are included for error analysis and future adjudication.

Why diagnostic metadata matters. Every crosswalk item should be distributed with a datasheet-style record: co-attested language count, positive prevalence in both databases, observed cosine, observed ϕ, Jaccard, predicted-profile cosine, nearest false positives, domain label, relation type, scope caveat, evidence type, semantic rationale, and a missingness policy. This is the crosswalk analogue of data statements and datasheets for NLP datasets (Bender and Friedman, 2018; Gebru et al., 2021): users need to know what a score permits them to do. A high observed-profile score is evidence for a retrieval candidate. A high learned-space cosine is weaker evidence: it suggests a typological hypothesis to inspect.

J Representative Crosswalk and Profile Diagnostics

Table 8 gives representative Tier-A rows. Accepted rows need not have uniformly high empirical agreement. A row can be semantically justified but empirically weak if one database codes dominance and the other codes presence, if the value is rare, if the two samples diverge, or if the relation is directional rather than symmetric.

Across the 76 Tier-A rows, median F1 is .644, median MCC is .449, and 60/76 rows have permutation $p \leq .05$. High-agreement rows are dominated by nominal-order correspondences. Weak rows are informative: WALS 59A counts possessive classifier classes, whereas GB058 asks whether possessive classifiers exist; WALS 37A distinguishes dedicated articles from demonstratives used as articles, whereas GB020 is broader in the definite/specific article domain while still excluding demonstratives that retain a clear spa-

WALS value	GB formula	Type	ϕ rank	Learned rank	MCC
89A=Noun–Numeral	GB024=2	Near	1	1	.875
89A=Numeral–Noun	GB024=1	Near	1	1	.824
88A=Demonstrative–Noun	GB025=1	Near	1	2	–
88A=Noun–Demonstrative	GB025=2	Near	1	1	.791
86A=Genitive–Noun	GB065=1	Near	1	5	.717
86A=Noun–Genitive	GB065=2	Near	1	3	–
87A=Adjective–Noun	GB193=1	Near	1	3	–
87A=Noun–Adjective	GB193=2	Near	1	9	–
85A=Prepositions	GB074=1	Dir.	1	2	.773
85A=Postpositions	GB075=1	Dir.	3	5	–
85A=No adpositions	GB074=0 & GB075=0	Near	1	1	–
37A=Definite article	GB020=1	Near	1	29	weak
38A=Indefinite article	GB021=1	Near	8	221	weak
59A=3–5 poss. classes	GB058=1	Dir.	–	–	-.082
118A=Nonverbal encoding	GB068=1	Near	–	–	-.582

Table 8: Representative Tier-A crosswalk rows and diagnostics. "Near" abbreviates NEAR-EXACT; "Dir." abbreviates DIRECTIONAL. Weak profile agreement is not automatically false; it flags scope, granularity, sample, prevalence, or polarity issues for manual inspection.

tial deictic function. Rows of this kind require a typed label and a caveat rather than an unqualified positive label.

K Cross-Database Retrieval and Confound Diagnostics

Table 9 gives the full denominator-matched retrieval table. The main paper reports the compact columns needed for the central argument; the appendix reports R@3 and median rank. The shared-denominator filter is applied before scoring: a Tier-A row is retained only if every method can assign a finite score. This removes four low-attestation rows affected by observed-profile coverage or neural vocabulary filtering. Using the same n for all methods is the fairest comparison because otherwise some methods are evaluated on easier or harder subsets.

The profile scores play different roles. Cosine, ϕ, and Jaccard are direct observed-profile anchors. Log-odds is included as a sparse-count robustness check, using add-.5 smoothing for zero cells; PPMI is included as a distributional-association baseline that is useful but more vulnerable to rare perfect co-occurrences. We therefore keep all ten methods in the compact main retrieval table, but reserve R@3, median rank, and scorer-specific interpretation for this appendix.

The genealogy baselines test whether retrieval is mainly family clustering. The updated denominator-matched results show that family-prevalence cosine is a strong confound baseline, especially at broad recall cutoffs. This makes the paper's caution stronger: high observed-profile agreement is most convincing when it exceeds family-prevalence and URIEL+ baselines and when the row's semantic rationale is independently auditable. URIEL+ phylogenetic centroids remain substantially lower than the best observed-profile anchors. The neural rows answer a different question: Learned predicted profiles retrieve more curated correspondences than T-CF under the same denominator, but neither should be used as the first-stage exact lookup signal.

L Within-Feature Substructure Protocol

The substructure probe is the main test of the categorical representation. If multi-valued features are contrast sets, then the value embeddings should recover internal relations among the alternatives: ordinal scales, absence-plus-subtype trees, Boolean or component lattices, permutation cross-products, and strategy families.

Diagnostic taxonomy. We annotate 70 WALS features with one or more substructure schemas. Ordinal features include size or count scales such as consonant inventory size (1A), inflectional synthesis (22A), number of genders (30A), number of cases (49A), and colour-category count (132A/133A). Presence-subtype trees include absent vs. present-with-subtypes systems such as tone (13A), reduplication (27A), definite/indefinite articles (37A/38A), numeral classifiers (55A), evidentiality (77A/78A), and antipassive (108A). Boolean/component lattices include combinations of marked components, such as glottalized consonants (7A), verbal person marking (102A), and asymmetric negation (114A). Permutation/cross-product systems include prohibitives (71A), SOV order (81A), object-oblique-verb order (84A), adposition order (85A), and OV crossed with adposition, relative-clause, or

Method	R@1	R@3	R@5	R@10	MRR	Med.	n
Observed profile anchors and count baselines							
Observed ϕ	.557	.729	.771	.814	.653	1	70
Observed Jaccard	.614	.714	.757	.771	.676	1	70
Observed cosine	.571	.686	.729	.771	.642	1	70
Log-odds	.443	.643	.714	.771	.555	2	70
PPMI	.157	.314	.486	.629	.299	6	70
Genealogy and phylogeny proxies							
Family-prevalence cosine	.500	.657	.671	.757	.586	1	70
URIEL+ phylogenetic centroid	.257	.400	.486	.514	.351	9	70
Neural predicted profiles and direct rotation							
Learned	.100	.200	.300	.371	.185	32	70
T-CF	.014	.029	.071	.100	.045	116	70
Procrustes A	.000	.000	.000	.000	.007	170	70

Table 9: Cross-database retrieval and confound diagnostics on the 70-pair Tier-A subset scored by all methods. Every method is evaluated on the same denominator. Observed and count-based rows use CLDF co-attestation profiles over 1,015 shared Glottocodes; neural rows average predicted-profile scores over five independently trained runs.

adjective order (95A–97A). Strategy-family features include alignment, negative morphemes, polar questions, predicative possession, and comparison strategies.

Metrics. Each schema receives a metric matched to its linguistic structure: Spearman's ρ between embedding distance and gold rank distance for ordinal scales; triplet accuracy for presence-subtype trees; leave-one-out ridge R^2 from component vectors for Boolean lattices; centroid-direction probe accuracy for permutation cross-products; and nearest-neighbour purity for strategy families. Diagnostics are computed over within-feature-centred embeddings: for each feature j, the mean of its K_j value embeddings is subtracted before cosines are computed. This removes feature-level offsets and isolates relations among values, matching the softmax objective's invariance to per-feature shifts.

Null baselines. Every diagnostic is compared against shuffled labels within the feature, random Gaussian embeddings of matched dimensionality, and a co-occurrence baseline that replaces learned value embeddings with binary language-incidence vectors over WALS-attested languages. The co-occurrence baseline is necessary because raw incidence profiles may already contain some of the target structure.

What the substructure results show. Ordinal features show the clearest architectural dissociation: T-CF has no reason to place values on a line and often inverts the gold rank direction. Presence-subtype trees are recovered reliably, with the largest article-system gap on 37A. Permutation cross-products reveal axis structure, especially in 84A verb position and 71A prohibitive. Boolean lattices are harder because the number of values is small relative to the component space, but Learned still improves consistently. Strategy-family nearest-neighbour purity is intentionally not interpreted as simple clustering: ergative and accusative values are competitors within a feature, but ergative values should find correlates across feature boundaries.

Known failures. Four features are uninformative or fail in the ordinal probe: 12A syllable structure, 39A inclusive/exclusive in independent pronouns, 66A past-tense remoteness, and 133A colour-category count. These are typically three-value features whose distributional typology is strongly confounded by macroarea. Colour-category systems, for instance, concentrate differently across Europe, East Asia, Africa, and Oceania. In these cases both learned architectures and the co-occurrence baseline fail, suggesting a typological sampling/coverage failure mode rather than an architecture-specific bug.

M Within-Feature Structural Projections and Distance Matrices

Figure 1 is the central main-paper visualization because it directly supports the title claim: multi-valued features have internal structure, and categorical supervision is better aligned with that structure than independent binary expansion. The main paper shows the distance matrices directly. This appendix gives the corresponding structural pro-

Feature / ordinal scale	Learned	T-CF	Cooc
1A Consonant inventory	+.68	+.05	-.53
3A Consonant-vowel ratio	+.42	-.19	-.53
22A Inflectional synthesis	+.52	-.45	-.64
26A Prefixing vs. suffixing	+.61	+.46	+.25
30A Number of genders	+.74	+.30	+.37
40A Inclusive/exclusive in verbs	+.41	-.59	+.12
49A Number of cases	+.48	+.33	+.25
53A Ordinal numerals	+.64	+.15	-.37
125A–128A Balanced/deranked	+.87	+.52	-.87

Table 10: Within-feature Spearman ρ between value-embedding cosine distance and gold ordinal rank distance. Learned recovers ordinal direction and is strictly better than T-CF on all representative ordinal rows.

Diagnostic	Feature(s)	Learned	T-CF	Cooc
Presence $\rightarrow$ subtype trees (triplet accuracy)				
Triplet	13A Tone	$1.00 \pm .00$	$1.00 \pm .00$	1.00
Triplet	37A Definite article	$1.00 \pm .00$	$.25 \pm .00$	.75
Triplet	38A Indefinite article	$1.00 \pm .00$	$.60 \pm .33$	1.00
Triplet	55A Classifiers	$1.00 \pm .00$	$1.00 \pm .00$	1.00
Triplet	77A Evidentiality	$1.00 \pm .00$	$.80 \pm .24$	1.00
Triplet	78A Evidential coding	$1.00 \pm .00$	$.60 \pm .25$	1.00
Boolean lattices (leave-one-out ridge R^2)				
R^2	4A Voicing	$-.57 \pm .06$	$-.88 \pm .44$	-1.35
R^2	102A Person marking	$-.82 \pm .06$	$-2.30 \pm .45$	-1.35
R^2	114A Asym. negation	$-.70 \pm .05$	$-.78 \pm .35$	-.95
Permutation/cross-product systems (probe accuracy)				
Probe	71A Imperative axis	$.90 \pm .12$	$.75 \pm .00$	.75
Probe	71A Negation axis	$1.00 \pm .00$	$.75 \pm .00$	1.00
Probe	84A Verb position	$1.00 \pm .00$	$.55 \pm .10$	.75
Probe	84A OV/VO	$1.00 \pm .00$	$1.00 \pm .00$	1.00
Probe	88A Direction	$1.00 \pm .00$	$.67 \pm .30$	.67
Probe	97A OV $\times$ adjective	$1.00 \pm .00$	$.60 \pm .20$	.75
Strategy families (within-feature nearest-neighbour purity)				
NN	98A Alignment	$.00 \pm .00$	$.00 \pm .00$	.00
NN	100A Verbal person	$.00 \pm .00$	$.00 \pm .00$	.00
NN	117A Predicative possession	$.00 \pm .00$	$.00 \pm .00$	.00

Table 11: Representative within-feature substructure results across diagnostic types, averaged over five independently trained runs. Negative Boolean-lattice R^2 is consistent with underpowered leave-one-out ridge settings with small K_j and should not be overinterpreted; Learned is nevertheless consistently less negative than T-CF and co-occurrence. Strategy-family purity of .000 is not automatically failure: strategy values can be paradigmatic competitors within a feature while correlating across feature boundaries.

jections and the full combined figure.

The three displayed features were chosen because they make distinct substructure types visible without requiring a large table. For 22A, Learned reproduces the ordinal staircase: distances grow with rank difference and run variance is small relative to between-cell differences. T-CF instead places the extreme values 2–3 and 8–9 categories per word close together. For 37A, Learned separates absence from definite-present subtypes, while T-CF collapses several non-identical article values. For 71A, Learned recovers the 2-by-2 prohibitive cross-product more clearly than T-CF. Additional matrix-style diagnostics for 81A word order, 102A person marking, and the remaining annotated features are included in the supplementary output rather than duplicated in the main PDF.

N Cross-Feature Neighbourhoods, Family Preservation, and Run Stability

Within-feature softmax turns values of the same feature into paradigmatic competitors. Positive typological affinity therefore often appears across feature boundaries rather than within them. In WALS Learned, 81A=SOV is nearest to 95A=OV+Postpositions (.984), 83A=OV (.975), 85A=Postpositions (.939), 144Q=No SNegOV (.860), and 86A=Genitive–Noun (.845), recovering the Greenbergian/Dryer head-final cluster (Greenberg, 1963; Dryer, 1992). Symmetrically, SVO neighbours include VOX, VO+NRel, VO+NAdj, and VO. Complex tone is close to isolating and zero-marking features often associated with a Southeast Asian areal profile, though

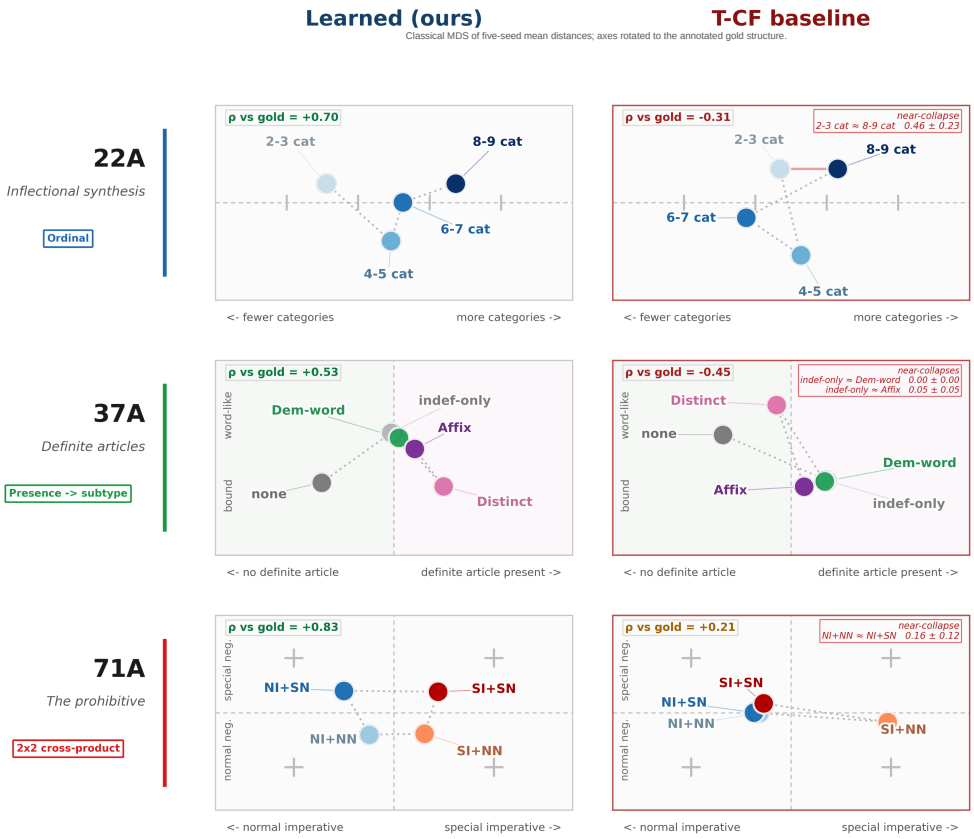

Figure 2: Projection half of Figure 1. Each row projects the five-run mean value-distance matrix for one WALS feature with classical MDS, then rotates the projection to the annotated gold structure. The projection view summarizes the same pairwise distances shown in the main paper: Learned preserves positive alignment with the ordinal scale in 22A, separates absent from present article values in 37A, and keeps both axes of the 71A prohibitive contrast more visible than T-CF.

this neighbourhood may also reflect genealogy, area, or coverage. Ergative-absolutive alignment is close across 98A and 99A and near oblique-patient constructions.

This also explains why strategy-family within-feature nearest-neighbour purity is .000 in several rows. Ergative and accusative alignment are alternatives inside 98A and should not be close simply because they belong to the same feature. The positive evidence is cross-feature: 98A=Ergative–absolutive is close to 99A=Ergative–absolutive and 108A=Oblique patient. The categorical architecture distinguishes alternatives of the same choice from correlates of the same typological profile; the binary baseline cannot represent that distinction directly.

Family preservation quantifies how much genealogical structure is encoded in the embedding space. Top-1 nearest-neighbour same-family rates over five runs are: WALS Learned .277±.004 (6.3× chance), WALS T-CF .125±.005 (2.9×), Grambank Learned .448±.003 (5.3×), and Grambank T-CF .230±.031 (2.7×), all $p < .002$ under permutation. Learned geometry is also more run-stable: Jaccard@10 over nearest-neighbour lists is .706 for WALS Learned and .672 for Grambank Learned, compared with .204 and .430 for T-CF. Genealogy should be read both as a quality signal and as a confound: for these database-derived embeddings, a typological space with no recoverable family structure is suspicious, but family structure alone can create false alignments.

O URIEL+/lang2vec Conditioning Sweep

The main text reports the substantive conditioning pattern; this appendix gives the full paired-evaluation design and the source-specific deltas. This experiment supplements the original branch-level prediction evaluation: each branch–fraction–repeat–conditioning cell is a paired comparison between the reused unconditioned F1 and

Database	Branches	ρ levels	Repeats	Variants	URIEL+ cov.	Result rows
WALS Learned	81	.01/.10/.20	5	geo/phylo/both	94.18%	7,290
Grambank Learned	52	.01/.10/.20	5	geo/phylo/both	94.62%	4,680

Table 12: Conditioning sweep design. Each result file contains paired baseline and conditioned rows, hence 2× branches × fractions × repeats × variants. The corresponding elapsed times are 84,761.8 seconds for WALS and 246,918.3 seconds for Grambank; the reported conditioning effects are aggregated over branch–fraction–repeat cells.

Database	Conditioning	Base	Cond.	Mean Δ	Median Δ	Win rate
WALS	geo	.5267	.5759	+.0492	+.0311	70.0%
WALS	phylo	.5267	.5682	+.0414	+.0281	68.7%
WALS	both	.5267	.5771	+.0503	+.0323	69.1%
Grambank	geo	.7096	.7139	+.0043	+.0059	61.5%
Grambank	phylo	.7096	.7121	+.0025	+.0056	61.3%
Grambank	both	.7096	.7137	+.0041	+.0046	57.6%

Table 13: Conditioning results by source, averaged over branches, supervision fractions, and repeats. Under this filtered branch-level setup, WALS shows a substantive gain; Grambank shows a small but mostly positive regularization effect.

a conditioned Learned model. The pipeline adds URIEL+/lang2vec geographic and phylogenetic vectors as residual offsets to the Learned model. Baseline F1 values are loaded from the pre-computed unconditioned comparison table, so each branch–fraction–repeat–conditioning cell trains only the conditioned model; the saved result rows nevertheless store both the baseline and conditioned F1 for paired comparison. This design follows the role of URIEL/lang2vec and URIEL+ as explicit sources of geographic and phylogenetic information rather than as hidden latent variables (Littell et al., 2017; Khan et al., 2025).

The conditioning results differ by resource and support a narrower claim. Under our split and filtering choices, WALS benefits most clearly from external geo/phylo priors where the prediction problem is most underdetermined: with only 1% in-branch supervision, all three conditioning variants give large positive deltas. Grambank, by contrast, is denser in our retained matrix and more genealogically structured under our diagnostics; its Learned space preserves family structure strongly, and external conditioning adds only a small regularizing effect. This helps explain why individual held-out branches can regress even when the aggregate effect is positive: the reported deltas are paired branch–fraction–repeat effects, and local regressions are expected when the external vector is misaligned with the held-out branch's observed typological profile.

P Universals Formula Probe

The universals probe uses a larger formula set derived from the GRAMBANK operationalizations used by Verkerk et al. (2026). That study reports tests of 191 proposed Universals Archive universals using Bayesian spatiophylogenetic regression and BayesTraits co-evolutionary analyses. We use their operationalized formulas and support labels as an external starting point, then apply our own coverage filter: after excluding formulas with sum(...) aggregates, no implication arrow, or values absent from the learned Grambank space, 172 universals remain testable.

For each formula $A \rightarrow C$, we resolve the antecedent and consequent into Grambank feature-value tokens. Simple formulas reduce to a single cosine. Compound formulas preserve logical structure: conjunctions and mixed antecedents use centroids; pure disjunctive antecedents use a conservative minimum over triggers; disjunctive consequents use a maximum; conjunctive consequents use a minimum; and mixed consequents fall back to a centroid. This yields M_1, a structured implication cosine. We also compute $M_2 = \cos(A, \neg C)$, an anti-cosine against the centroid of opposite consequent values, and an analogy residual $\|(A - \neg A) - (C - \neg C)\|$. Null distributions are cross-feature: 30,000 random token pairs for M_1 and 10,000 random quadruples for analogy residuals.

These results should not be interpreted as treat-

Database	ρ	geo Δ	geo win	phylo Δ	phylo win	both Δ	both win
WALS	.01	+.1145	86.9%	+.0884	78.3%	+.1190	87.4%
WALS	.10	+.0144	58.8%	+.0157	59.5%	+.0143	56.8%
WALS	.20	+.0187	64.4%	+.0202	68.4%	+.0176	63.5%
Grambank	.01	+.0002	50.0%	-.0049	45.4%	+.0025	48.8%
Grambank	.10	+.0050	65.8%	+.0044	65.8%	+.0033	56.9%
Grambank	.20	+.0078	68.8%	+.0081	72.7%	+.0067	66.9%

Table 14: Mean F1 deltas and win rates by in-branch supervision fraction. The WALS pattern is a cold-start-prior effect: gains are largest at $\rho = .01$ and shrink once in-branch evidence is available. The Grambank pattern is almost the reverse: conditioning is neutral or slightly harmful at $\rho = .01$ and becomes mildly positive at $\rho = .10$–$.20$.

Probe quantity	Value
Testable formulas	172 / 191
Mean M_1 over all formulas	+.357
Fully supported M_1	+.684
Unsupported M_1	+.185
Full − unsupported	+.499; MWU $p < 10^{-4}$
Full > BRMS-only	MWU $p = .0003$
Mean anti-cosine	-.415
Mean implication margin	+.772
Analogy-significant rows	19 (T-CF: 0)

Table 15: Universals formula probe using formulas derived from the Verkerk et al. Grambank operationalizations. M_1 is the structured implication cosine; values are five-run Learned means unless stated otherwise.

ing every co-occurrence as a universal. Instead, the Learned geometry is more consistent with the subset of formula-level implications that survive an external spatiophylogenetic and coevolutionary evaluation. Two boundary conditions remain. Coexistence universals are not well tested by value-vector similarity: a formula asking whether two strategies can co-occur is not the same as a formula asking whether their profiles are distributionally similar. Absence-conditioned antecedents are also unstable, because a value such as "not prefixal object indexing" pools heterogeneous language types. These cases may require a representation that models co-occurrence and absence more directly; we treat them as boundary cases rather than failures.

Q Projection and Subspace Diagnostics

The projection figures are used only as qualitative diagnostics. t-SNE and UMAP are visualization methods, not estimators of typological distance (van der Maaten and Hinton, 2008; McInnes et al., 2018). The quantitative claims in the main text are based on within-feature cosine distances, family-preservation rates, run stability, and Procrustes controls, not on visual projection. Figures 3–5 display a visual summary.

R Tree-Clustering Baseline

In our run, a ϕ-coefficient UPGMA clustering over the binary language-by-feature-value matrix recovers the head-final cluster as a single subtree. Cophenetic correlations are .359 for WALS and .490 for Grambank; VO–preposition has $\phi = +.80$ and OV–postposition has $\phi = +.79$. This baseline does not perform cross-database alignment. Cross-database profile similarity extends the same co-distributional intuition to the inter-database setting.

S Compute

GPU experiments used one A100-class GPU. Leave-one-branch-out factorial experiments used approximately 20 GPU-hours; canonical full-data models plus run-stability retraining approximately 10 GPU-hours; profile computation and alignment approximately 2 GPU-hours; substructure diagnostics approximately 0.1 GPU-hours because they are pure analysis over checkpoints; the earlier $\rho = .20$ conditioning pilot approximately 2 GPU-hours; the expanded Learned conditioning sweep 23.5 wall-clock hours for WALS and 68.6 for Grambank with baseline reuse; and direction-profile plus dependency-linearization runs approximately 2–4 GPU-hours depending on sentence cap and device. The full set of experiments used approximately 125–130 GPU-hours, dominated by the expanded conditioning sweep.

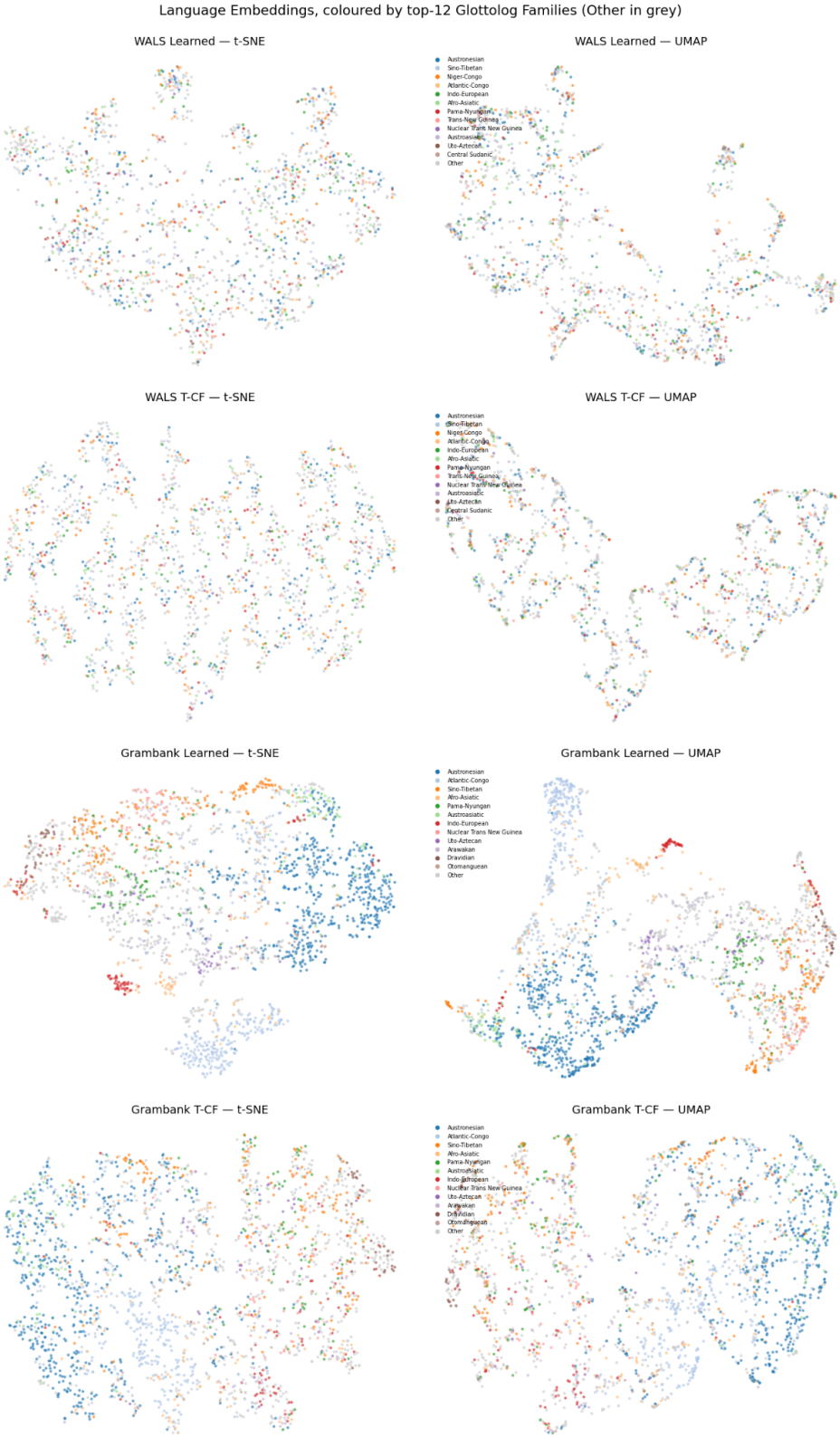

Figure 3: Language embeddings coloured by the top-12 Glottolog families. Grambank Learned exhibits the clearest family clustering; WALS spaces are more diffuse and T-CF is less stable.

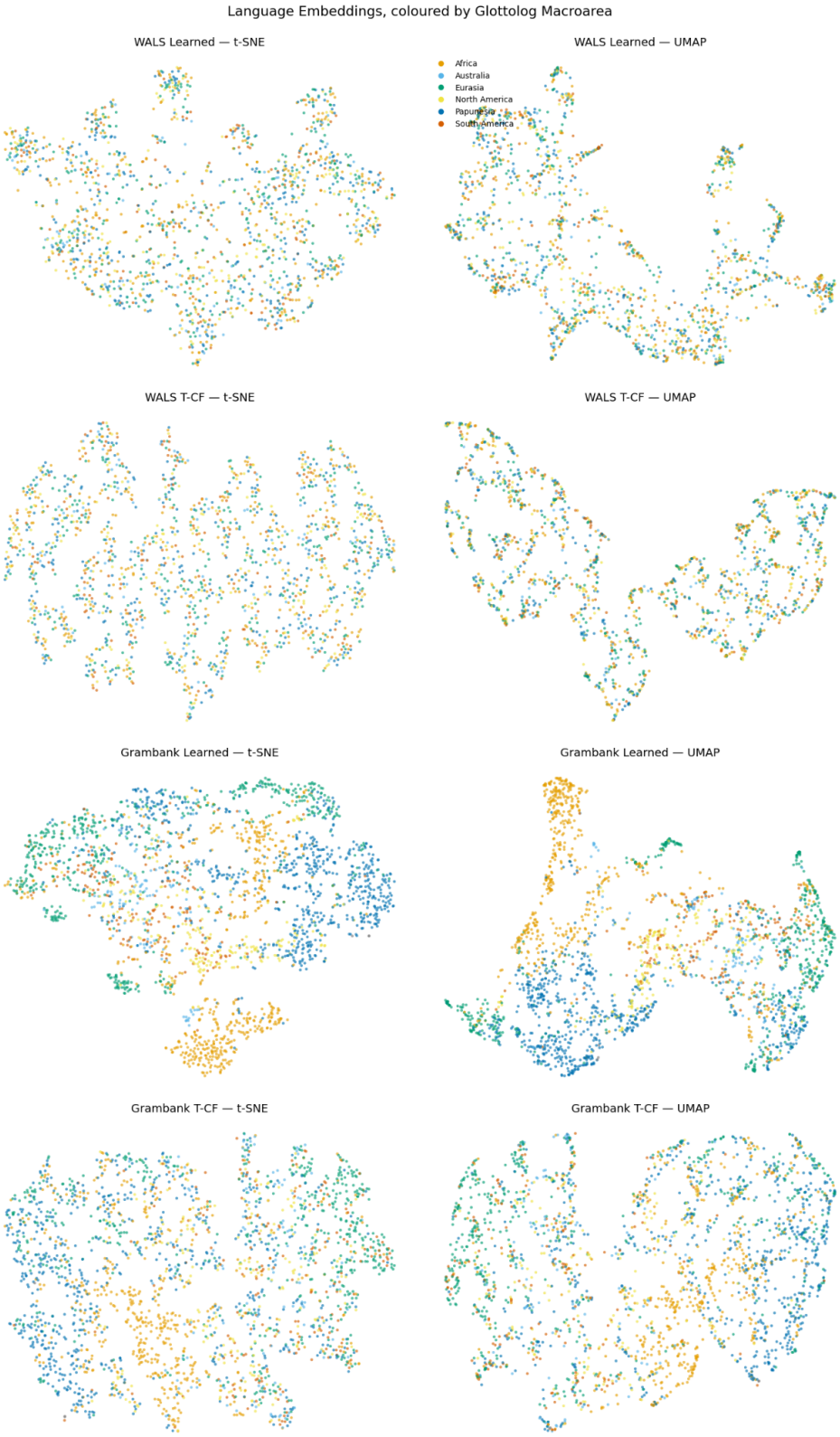

Figure 4: Language embeddings coloured by Glottolog macroarea. Macroarea signal is visible but entangled with genealogy and documentation coverage, motivating explicit geo/phylo conditioning rather than leaving this structure as an uninterpreted latent effect.

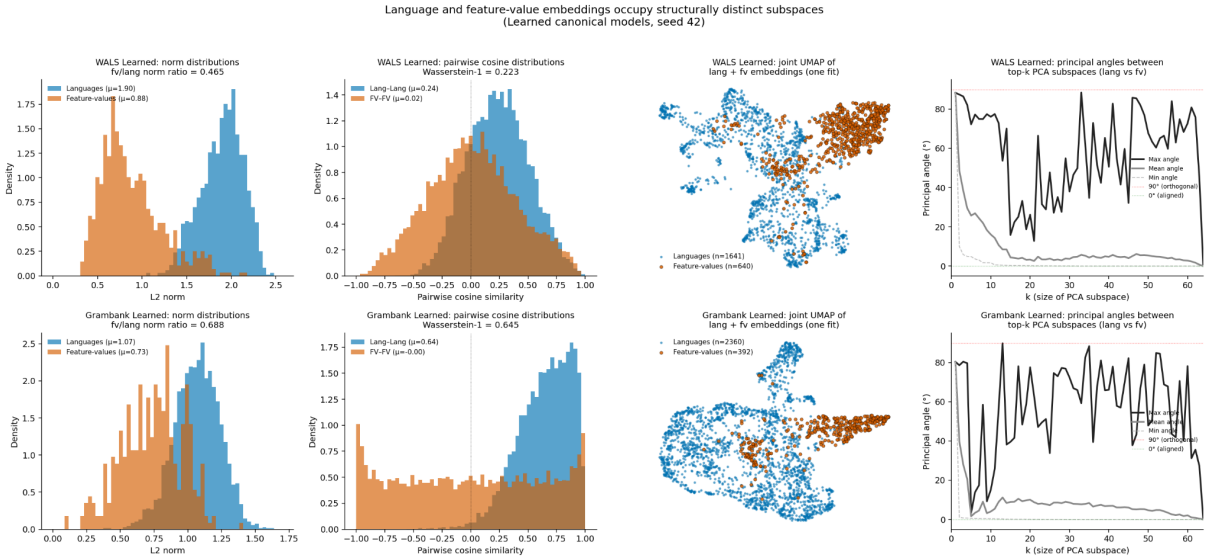

Figure 5: Language and feature-value embeddings occupy structurally distinct subspaces. Norm distributions, pairwise cosine distributions, joint UMAPs, and principal-angle diagnostics help diagnose why direct value-space Procrustes alignment fails. Profile-based alignment is therefore used for retrieval, while cosine distances among value embeddings are reserved for substructure analysis.