

Reading Relatedness and Hierarchy from Frozen Representations

Anonymous Authors

ANONYMOUS@OPENREVIEW.NET

Abstract

What does agreement between representation geometry and a reference hierarchy establish? We separate pairwise relatedness, agreement on individual four-source comparisons, and recovery of reference splits in a reconstructed tree. Across frozen text, speech, protein, vision and single-cell representations, these measurements support different conclusions. Local quartet agreement ranges from 0.530 to 0.854, but neighbor joining and FastME recover only a minority of the specified splits in the four larger populations. Nevertheless, their reconstructed trees retain substantial quartet agreement: low exact split recall does not imply absence of hierarchical information. Matched readout comparisons expose a second distinction. For 36 languages, simply centering raw centroids changes the cosine association with genealogy, adjusted for typology, from 0.053 to 0.432, approximately matching the aligned readout. Matching the origin also largely removes an apparent alignment-population effect. Exact additive-distance controls validate both reconstruction pipelines; a separate Brownian feature simulation illustrates how perturbations affect graded agreement and split recovery. The resulting protocol treats the origin, normalization, layer, aggregation and reconstruction method as parts of the measurement. It supports specific claims about accessible structure, without equating a failed readout with information absent from the representation.

Keywords: representation geometry, hierarchical structure, phylogenetic reconstruction, readout sensitivity, self-supervised learning

1. Introduction

Related sources often occupy nearby positions in a learned representation. This observation can support several progressively more specific interpretations: the representation distinguishes related from unrelated sources; it agrees with local branching relationships; or it permits reconstruction of a particular hierarchy. The measurements needed for these interpretations are different. A high nearest-neighbor score does not determine a tree, and a low rate of exact split recovery does not establish that a reconstructed tree is uninformative.

Consider the vision case in this study. For DINOv2 centroids of CIFAR-100 classes, neighbor joining recovers only 2 of 25 specified semantic splits using cosine dissimilarity. Yet the resulting tree agrees with 0.822 of the reference-resolved quartets. These are compatible descriptions of the same reconstruction. An exact split requires every source to fall on its correct side; quartet agreement measures how much local structure remains. The reference itself is a semantic taxonomy, so neither result is evidence about evolutionary descent.

The geometry being evaluated also requires specification. Changing the origin leaves Euclidean distances unchanged but can substantially change cosine dissimilarities. Source-specific alignment, radial normalization and layer selection introduce further choices. Consequently, a change in an association after preprocessing need not isolate the operation to

which it is attributed. We use matched comparisons to determine what these choices change in the cached representations, rather than select a readout by its agreement with the target.

Our contribution is an empirical methodology with three components. First, five case studies compare local relatedness and hierarchical recovery with a shared evaluator and two distance-based tree estimators. Second, matched language readouts and a speech layer sweep identify consequential measurement choices, including an origin effect that explains much of an apparent alignment effect. Third, exact-distance controls and a separate feature-space simulation distinguish estimator validation from an illustrative perturbation mechanism. Together they support a more precise question than whether a representation “contains a phylogeny”: *which reference relationships are accessible through a declared readout, at which resolution?*

This question connects representation probing to established work on distance-based reconstruction. Language vectors and multilingual encoders have been used to study genealogical and typological structure (Östling and Tiedemann, 2017; Bjerva et al., 2019; Rama et al., 2020). The distinction between a representation’s information and a probe’s accessibility is central to probe interpretation (Hewitt and Liang, 2019; Pimentel et al., 2020); structural probes explicitly learn a geometry for a target relation (Hewitt and Manning, 2019). Here we study specified distance readouts without training a supervised tree probe. Our comparisons also complement evidence that embedding geometry depends on influential dimensions and normalization (Timkey and van Schijndel, 2021).

2. A protocol for measuring accessible structure

Sources, representations and references. A source is a language, species, image class or cell type. Each source is reduced to a vector; Table 1 states the aggregation and the meaning of its reference hierarchy. The five populations are case studies with different sample sizes and reference resolutions. They are not exchangeable replications of one evolutionary process; text and speech also share language histories. We retain the submitted cross-domain readouts and assess alternatives in explicit sensitivity analyses. Extraction details and all reference paths are recorded in the artifact and Appendix A.

Only the text experiment uses Procrustes alignment. Its rotations and per-language scales are fitted on 812 parallel sentences; centroids use the remaining 200. The stored origin is a grand mean that includes both portions, making the origin transductive. Speech, vision and cell vectors are pooled in the encoder’s coordinates. Protein vectors concatenate coordinate-standardized gene embeddings, with missing gene blocks set to zero. These reductions discard within-source variation; all claims below concern the resulting vectors, not every statistic of the original representation clouds.

Three resolutions of agreement. For source vectors x_i , we evaluate Euclidean distance and cosine dissimilarity $D_{ij}^C = 1 - x_i^\top x_j / (\|x_i\| \|x_j\|)$. We call the latter a dissimilarity because it need not satisfy the triangle inequality. The reference matrix G is the path distance in the coded hierarchy with unit edges; it is not a matrix of divergence dates. We report:

1. **Pairwise relatedness:** Pearson correlation r between the upper triangles of D and G , and nearest-neighbor group purity. Purity is averaged over tied nearest neighbors and compared with $\sum_g n_g(n_g - 1) / [n(n - 1)]$ for the stated grouping.

Domain	Frozen representation / aggregation	Reference interpretation	Sources	Splits
Text	XLM-R-large, aligned sentence centroid	Coded family/branch/genus hierarchy	89	36
Speech	XLS-R-300m, utterance centroid	Coded language family hierarchy	30	7
Protein	ESM-2-650M, concatenated gene vectors	Coarse vertebrate class/subgroup hierarchy	25	6
Vision	DINOv2-base, image-class centroid	CIFAR-100 superclasses plus five authored metaclasses	100	25
Single-cell	Geneformer V2-104M, cell-type centroid	Lymphoid/myeloid grouping of PBMC types	8	2

Table 1: The populations differ in what their references assert. Splits are unique nontrivial unrooted bipartitions; none of these empirical references fully resolves every source relationship. The cell grouping is not a measured developmental lineage tree.

2. **Local quartet agreement** Q_D : for each reference-resolved set $\{i, j, k, l\}$, compare the minimizing pairing among $D_{ij} + D_{kl}$, $D_{ik} + D_{jl}$ and $D_{il} + D_{jk}$ with the reference pairing. This is a local topological statistic, not a purely pairwise metric statistic.
3. **Reconstructed hierarchy**: infer an unrooted tree, then report both reference-split recall $S = |\mathcal{S}(\hat{T}) \cap \mathcal{S}(G)|/|\mathcal{S}(G)|$ and quartet agreement $Q_{\hat{T}}$ of that tree. A split and its complement count once; both sides must contain at least two sources. Tree quartets use topology with unit edges, independently of estimated branch lengths.

The empirical references contain unresolved relationships. We therefore score only quartets that the reference resolves and do not label split recall “exact tree recovery.” In particular, we do not penalize an inferred tree merely for resolving a reference polytomy. Full-tree exactness and normalized Robinson–Foulds distance are additionally recorded for the fully resolved synthetic trees. Predicted quartet ties receive fractional credit; 1/3 is the uniform-guess baseline.

Estimators and uncertainty. We compare neighbor joining (NJ; [Saitou and Nei 1987](#)) with FastME 2.1.6.3 balanced minimum evolution, using balanced taxon addition and its NNI and SPR searches ([Lefort et al., 2015](#)). FastME broadens the reconstruction comparison, but both procedures act on the same distance matrices and share a minimum-evolution motivation. Their agreement cannot establish a limit for other estimators or representation statistics.

We enumerate all quartets for populations of at most 30 sources and otherwise sample 40,000 distinct quartets uniformly, sharing the sample across readouts and methods. Unadjusted r , Q_D and purity use 2,000 source-label permutations, with $p = (h + 1)/(2001)$; r is two-sided and agreement statistics upper-tailed. These test a random-label null, not a causal account of relatedness. Pairs and quartets sharing sources are not independent replicates. Adjusted language correlations are descriptive residual correlations, accompanied by leave-one-source-out sensitivity ranges rather than conditional-Mantel significance

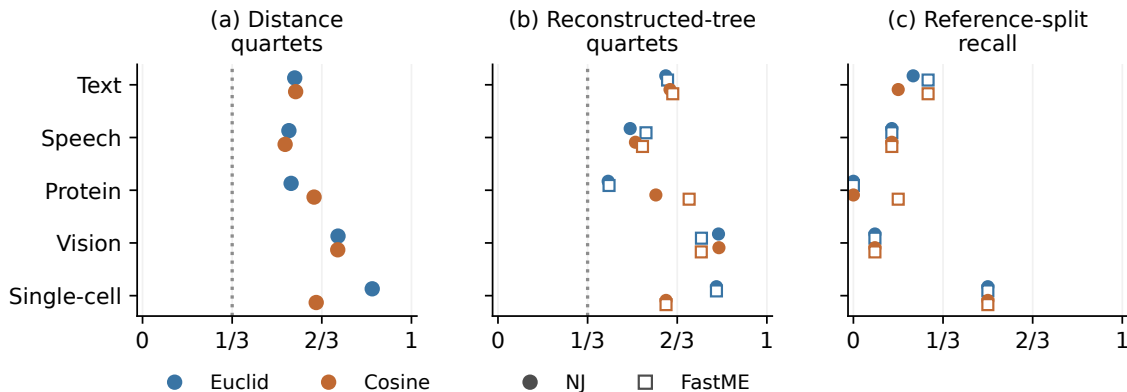


Figure 1: Three resolutions of agreement, evaluated on the same sources and reference hierarchies. Color distinguishes Euclidean and cosine readouts; filled circles and open squares distinguish NJ and FastME in (b,c). The dotted line in (a,b) is the uniform quartet-guess baseline, not a confidence threshold. Split recall in (c) has no shared chance baseline across these differently resolved references. Complete counts and random-label tests appear in Appendix B.

claims (Guillot and Rousset, 2013). Cached centroids do not permit an observation-level bootstrap. The audit is retrospective; its protocol and controls were specified before the new estimator comparisons were run.

3. Local agreement and exact splits give different readings

Local quartet agreement exceeds $1/3$ in all ten domain/readout combinations, ranging from 0.530 to 0.854 (Figure 1a). Nearest-neighbor purity also exceeds its group-size baseline throughout. This does not imply uniformly strong distance-matrix correlation: speech has $r = 0.137$ under Euclidean distance, with random-label $p = 0.1144$; the Euclidean protein correlation is also weak. The phrase “faithful distances” would therefore be too broad even before considering reconstruction.

Exact reference-split recall is low for both reconstruction methods in the four larger populations, but it is method-dependent. For text, NJ recovers $8/36$ splits with Euclidean distance and $6/36$ with cosine; FastME recovers $10/36$ under each readout. Protein changes from no recovered splits with NJ to $1/6$ with cosine FastME. Speech recovers one of seven and vision two of 25 under both methods and metrics. The cell case recovers one of only two specified splits. It provides a small illustration, not a precise estimate of recovery on developmental trees.

The reconstructed trees nevertheless preserve graded hierarchical agreement (Figure 1b). Vision reaches $Q_{NJ} = 0.822$ under cosine even though its split recall is 0.080. For protein, cosine FastME reaches $Q_{FM} = 0.711$, compared with $Q_{NJ} = 0.587$. Conversely, FastME reduces vision quartet agreement while retaining the same split count. Optimizing the reconstruction objective need not improve every reference-based score. These results support

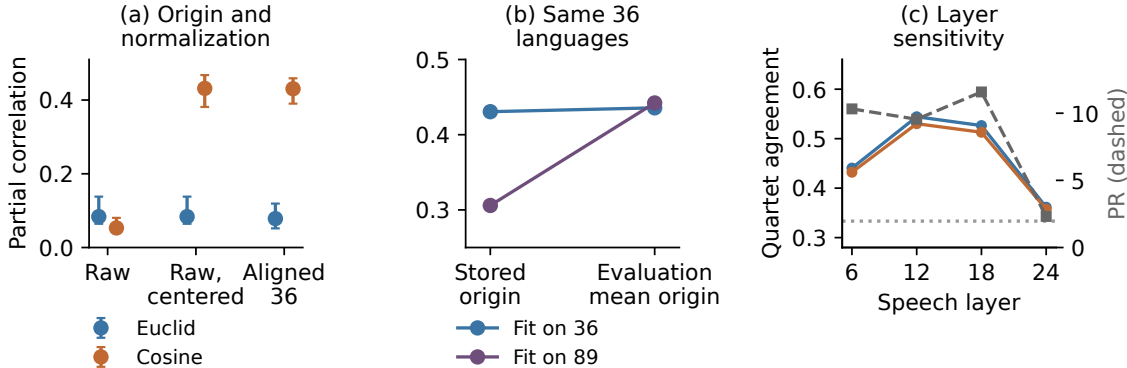


Figure 2: Matched readout comparisons. (a) The same 36 languages and layer: changing the origin of raw centroids substantially changes the adjusted cosine association. Bars span source-deletion estimates; they are sensitivity ranges, not confidence intervals. (b) Evaluating the same 36 languages after fitting on 36 or 89: subtracting each set’s evaluation-centroid mean largely closes the gap. (c) Speech quartet agreement and participation ratio vary across cached layers. The dashed PR curve uses the right axis; the horizontal dotted line marks quartet chance.

limited exact split recovery under the evaluated pipelines, together with substantial partial recovery; they do not support absence of phylogenetic information.

Reference choice also affects apparently simple comparisons. Vision’s cosine nearest-neighbor purity is 0.960 for the five authored metaclasses, but 0.680 for the 20 official CIFAR superclasses, whose chance level is 0.040. Protein uses a coarse taxonomic grouping and uneven gene coverage, rather than a dated molecular phylogeny. These distinctions limit biological interpretation and make a pooled cross-domain recovery rate inappropriate.

4. Readout choices can explain apparent structural effects

The origin precedes the radial quotient. On a 36-language subset, let G_U be URIEL family-vector distance and T its imputed syntactic distance (Littell et al., 2017). Unlike the coded hierarchy in Section 3, these are feature-vector distances. We compare $\text{corr}(D, G_U | T)$ using a fixed pair set. At the stored text layer L12, raw Euclidean and cosine readouts yield 0.084 and 0.053, respectively. Subtracting the mean of these 36 raw centroids leaves Euclidean distances unchanged, but raises the cosine association to 0.432. The stored alignment fitted on these same languages yields 0.431. Thus the large adjusted association does not require the source-specific alignment in this comparison. All four cached layers and both evaluation populations are reported in Appendix C; L12 is an illustration, not a layer selected by a new optimization.

For an explicit origin o , cosine compares the directions of $x_i - o$ and $x_j - o$. It removes positive rescaling of these vectors, but it is not translation-invariant. Euclidean distance obeys

$$\|x_i - x_j\|^2 = \|x_i - o\|^2 + \|x_j - o\|^2 - 2\|x_i - o\|\|x_j - o\|\cos\theta_{ij}(o),$$

so its radial and angular decomposition itself depends on α . In the 89-fitted frame restricted to the 36 languages, L12 norm correlates with tokenizer fragmentation (subwords per character) at 0.824. This association motivates checking radial sensitivity. It does not prove that the radial component contains only nuisance or that removing it recovers a uniquely correct geometry. Norms can carry useful information in other settings (Oyama et al., 2023).

Matching the origin changes the alignment-population interpretation. For those same 36 languages, cosine partial correlation is 0.431 when alignment is fitted on 36 and 0.306 when their vectors are taken from the 89-language fit. This comparison changes both the fitted transformations and the origin. After subtracting the mean of the evaluated 36 centroids within each frame, the values become 0.436 and 0.442. The original gap therefore cannot be attributed to the alignment population independently of centering. This matched-origin operation is label-free but transductive: it uses the evaluation population to define its origin, and is not an independently validated rule for future sources.

Effective dimension diagnoses a readout, not its information capacity. For centered source vectors with singular values s_j , participation ratio is $\text{PR} = (\sum_j s_j^2)^2 / \sum_j s_j^4$. The final speech layer has PR 2.315, out of a maximum centered rank of 29, and mean pair-wise cosine 0.998. Its cosine quartet agreement is 0.357, compared with 0.530 at hidden-state index 12, whose PR is 9.549. PR flags concentration of the measured variation; it neither proves that no hierarchy is encoded nor identifies the best layer without evaluation. Layer sweeps should accompany such diagnostics.

Matching pairs changes the control comparison. A similar attribution issue arises in the language typology control. Within six families, an eight-configuration aligned Euclidean ensemble yields an adjusted genealogical correlation of 0.264 with imputed typology over 94 pairs, versus 0.349 with co-observed typological features over 81 pairs. On the *same* 81 pairs, the imputed-control result is 0.352. The unmatched contrast therefore provides no evidence for a large imputation penalty. The retained descriptive association is useful, but its interpretation requires specifying both feature provenance and pair availability (Appendix D).

5. Validation and a controlled perturbation example

An exact-distance control. We first give the estimators the exact path-distance matrices of balanced binary trees with 8 and 256 leaves. Both NJ and FastME recover every unique reference split and every tested quartet. These controls verify the distance-to-tree and scoring pipelines under exact additivity. They do not establish that either estimator is optimal on the empirical dissimilarities.

A separate feature-space simulation. We then generate 64-dimensional leaf vectors by summing independent $\mathcal{N}(0, I_{64})$ edge increments on a balanced depth-8 tree. If d_T is its unit-edge distance, then

$$\mathbb{E}[\|Z_i - Z_j\|^2/64] = d_T(i, j).$$

The observed finite-dimensional Euclidean distances are *not* exact additive tree distances; neither are their cosine counterparts. This is a noisy positive example distinct from the

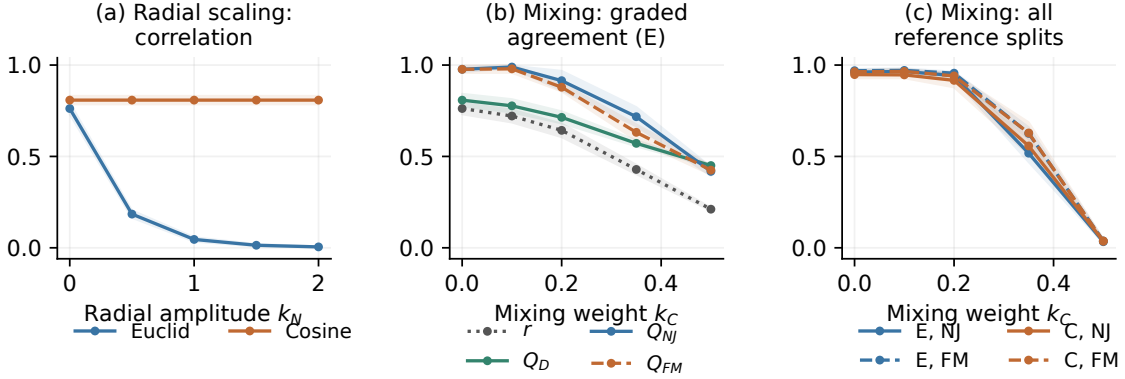


Figure 3: Perturbations of Brownian leaf features, separate from the exact-distance control. Lines are means over five seeds; bands show one seed standard deviation. (a) Positive radial scaling preserves cosine dissimilarities exactly. (b) Donor mixing changes distance correlation, local quartet agreement and reconstructed-tree quartet agreement at different rates; E denotes Euclidean distance. (c) Recall of all 253 unique reference splits under both readouts and estimators. The full numerical sweep and deep-split scores appear in Appendix E.

oracle control. We evaluate all 253 unique internal splits with the same scorer used above, and separately report the five unique splits induced by the two deepest levels.

We apply two feature-space perturbations: a positive radial scale $e^{k_N T_i}$, with fixed $r_i \sim \mathcal{N}(0, 1)$, and donor mixing $Z_i \mapsto (1 - k_C)Z_i + k_C Z_{d(i)}$, with a fixed donor from the opposite root subtree. We vary one amplitude at a time, sharing the tree, scales, donors and quartet sample across doses within each of five seeds. This interpolation is a sufficient construction for studying altered geometry; it is not a model fitted to contact, horizontal transfer or encoder training.

Without perturbation, mean Euclidean NJ split recall is 0.961, while FastME yields 0.969. Positive radial scaling reduces Euclidean correlation but leaves the cosine measurements unchanged, as required algebraically. At mixing weight $k_C = 0.35$, Euclidean distance correlation is 0.429, local quartet agreement is 0.572, and all-split recall is 0.518 for NJ and 0.626 for FastME. At this dose neither method recovers any of the five deepest splits, even while many of the remaining splits survive. Thus the apparent ordering of “topology loss” depends on which splits are scored. Reporting the full hierarchy avoids reducing the response to a selected success/failure boundary.

What the recovery guarantee explains. For an additive binary-tree distance d_T with minimum edge length $b > 0$, Atteson’s sufficient condition for NJ recovery is $\|\hat{d} - d_T\|_\infty < b/2$ (Atteson, 1999). Exceeding this radius removes that guarantee; it does not require NJ to fail. NJ can succeed outside it (Mihaescu et al., 2009). The condition concerns worst-case distance error relative to edge length, whereas a global correlation measures average variation across pairs.

Under the additional idealized model $\hat{d} = d + \varepsilon\eta$, treating upper-triangle entries as the variables and assuming $\text{Cov}(d, \eta) = 0$, nonzero $\text{Var}(d)$ and finite second moments,

$$\text{corr}(d, \hat{d}) = \left(1 + \varepsilon^2 \frac{\text{Var}(\eta)}{\text{Var}(d)}\right)^{-1/2} = 1 - \frac{\varepsilon^2}{2} \frac{\text{Var}(\eta)}{\text{Var}(d)} + O(\varepsilon^4).$$

This separates a global correlation scale from an exact-recovery guarantee. The empirical embeddings and the donor-mixing experiment are not assumed to satisfy this noise model. Thus the theorem, the illustrative simulation and the empirical observations provide different kinds of evidence; none identifies the cause of reconstruction errors in the real populations.

6. Discussion and limits

The strongest supported conclusion is that accessible structure has both a *resolution* and a *readout*. Local agreement can coexist with low exact split recall, while reconstructed trees still preserve many reference quartets. Meanwhile, apparently substantive differences between readouts can arise from unmatched origins or pair sets. A useful reporting protocol therefore declares the source aggregation, origin, normalization, layer and fitted population; evaluates relatedness, local quartets and reconstructed-tree agreement separately; validates the scorer on known trees; and compares proposed explanations under matched conditions.

These checks do not select a universally correct geometry. A common orthogonal transformation preserves norms, angles and Euclidean distances alike, so orthogonal invariance does not uniquely privilege cosine. Source-specific rotations, scaling and centering are different operations. Identifiability results for specified contrastive objectives and generative models (Zimmermann et al., 2021) do not by themselves justify an invariant readout across the heterogeneous pretrained encoders studied here. Label-free diagnostics such as PR, radial associations and origin sensitivity can reveal fragile measurements; target agreement is still needed to assess what a readout recovers.

The experiments also have substantive limits. The empirical references are coarse and partly authored. Protein retrieval uses gene-name queries rather than verified orthology, with strongly uneven coverage and zero-filled missing blocks. The cell population contains only eight types and two scored groups. Text uses parallel translated sentences, so corpus and translation effects may contribute to the measured associations. Cloud aggregation excludes covariance, multimodality and subspace-specific structure. Archived language cloud-statistic results are quantified in Appendix E.1, but do not exclude richer distributional readouts.

Likewise, neither a small *mean* four-point residual nor its diameter normalization rules out local distortion at short branches. We make no exclusion claim about hyperbolic readouts or alternative tree embeddings, and do not infer a training-objective mechanism without a controlled training intervention. Nonlinear distance transformations and learned probes remain possible alternatives. The present results justify precise statements about the tested measurements: partial hierarchical structure is accessible, exact reference splits are recovered unevenly, and the preprocessing that defines the geometry materially changes the answer.

References

- Kevin Atteson. The performance of neighbor-joining methods of phylogenetic reconstruction. *Algorithmica*, 25(2–3):251–278, 1999. doi: 10.1007/PL00008277.
- Arun Babu, Changhan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick von Platen, Yatharth Saraf, Juan Pino, Alexei Baevski, Alexis Conneau, and Michael Auli. XLS-R: Self-supervised cross-lingual speech representation learning at scale. In *Interspeech 2022*, pages 2278–2282, 2022.
- Johannes Bjerva, Robert Östling, Maria Han Veiga, Jörg Tiedemann, and Isabelle Augenstein. What do language representations really represent? *Computational Linguistics*, 45(2):381–389, 2019. doi: 10.1162/coli_a.00351.
- Will Chang, Chundra Cathcart, David Hall, and Andrew Garrett. Ancestry-constrained phylogenetic analysis supports the Indo-European steppe hypothesis. *Language*, 91(1):194–244, 2015.
- Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.acl-main.747.
- Alexis Conneau, Min Ma, Simran Khanuja, Yu Zhang, Vera Axelrod, Siddharth Dalmia, Jason Riesa, Clara Rivera, and Ankur Bapna. FLEURS: Few-shot learning evaluation of universal representations of speech. In *IEEE Spoken Language Technology Workshop (SLT)*, 2022.
- Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. Language-agnostic BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891. Association for Computational Linguistics, 2022.
- Gilles Guillot and François Rousset. Dismantling the mantel tests. *Methods in Ecology and Evolution*, 4(4):336–344, 2013. doi: 10.1111/2041-210X.12018.
- John Hewitt and Percy Liang. Designing and interpreting probes with control tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2733–2743. Association for Computational Linguistics, 2019. doi: 10.18653/v1/D19-1275.
- John Hewitt and Christopher D. Manning. A structural probe for finding syntax in word representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*, pages 4129–4138. Association for Computational Linguistics, 2019. doi: 10.18653/v1/N19-1419.

- Eric W. Holman, Cecil H. Brown, Søren Wichmann, André Müller, Viveka Velupillai, Harald Hammarström, Sebastian Sauppe, Hagen Jung, Dik Bakker, Pamela Brown, Oleg Belyaev, Matthias Urban, Robert Mailhammer, Johann-Mattis List, and Dmitry Egorov. Automated dating of the world’s language families based on lexical similarity. *Current Anthropology*, 52(6):841–875, 2011. doi: 10.1086/662127.
- Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.
- Vincent Lefort, Richard Desper, and Olivier Gascuel. FastME 2.0: A comprehensive, accurate, and fast distance-based phylogeny inference program. *Molecular Biology and Evolution*, 32(10):2798–2800, 2015. doi: 10.1093/molbev/msv150.
- Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, Allan dos Santos Costa, Maryam Fazel-Zarandi, Tom Sercu, Salvatore Candido, and Alexander Rives. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, 2023. doi: 10.1126/science.ade2574.
- Patrick Littell, David R. Mortensen, Ke Lin, Katherine Kairis, Carlisle Turner, and Lori Levin. 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: Volume 2, Short Papers*, pages 8–14. Association for Computational Linguistics, 2017.
- Radu Mihaescu, Dan Levy, and Lior Pachter. Why neighbor-joining works. *Algorithmica*, 54(1):1–24, 2009. doi: 10.1007/s00453-007-9116-4.
- NLLB Team. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*, 2022.
- Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. DINOv2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.
- Robert Östling and Jörg Tiedemann. Continuous multilinguality with language vectors. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 644–649. Association for Computational Linguistics, 2017.
- Momose Oyama, Sho Yokoi, and Hidetoshi Shimodaira. Norm of word embedding encodes information gain. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2108–2130. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.emnlp-main.131.
- Tiago Pimentel, Josef Valvoda, Rowan Hall Maudslay, Ran Zmigrod, Adina Williams, and Ryan Cotterell. Information-theoretic probing for linguistic structure. In *Proceedings of*

- the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4609–4622. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.acl-main.420.
- Taraka Rama, Lisa Beinborn, and Steffen Eger. Probing multilingual BERT for genetic and typological signals. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1214–1228. International Committee on Computational Linguistics, 2020. doi: 10.18653/v1/2020.coling-main.105.
- Naruya Saitou and Masatoshi Nei. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution*, 4(4):406–425, 1987. doi: 10.1093/oxfordjournals.molbev.a040454.
- Alexander Savelyev and Martine Robbeets. Bayesian phylolinguistics infers the internal structure and the time-depth of the Turkic language family. *Journal of Language Evolution*, 5(1):39–53, 2020.
- Christina V. Theodoris, Ling Xiao, Anant Chopra, Mark D. Chaffin, Zeina R. Al Sayed, Matthew C. Hill, Helene Mantineo, Elizabeth M. Brydon, Zexian Zeng, X. Shirley Liu, and Patrick T. Ellinor. Transfer learning enables predictions in network biology. *Nature*, 618(7965):616–624, 2023. doi: 10.1038/s41586-023-06139-9.
- William Timkey and Marten van Schijndel. All bark and no bite: Rogue dimensions in transformer language models obscure representational quality. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4527–4546. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.emnlp-main.372.
- Grace X. Y. Zheng, Jessica M. Terry, Phillip Belgrader, Paul Ryvkin, Zachary W. Bent, Ryan Wilson, Solongo B. Ziraldo, Tobias D. Wheeler, Geoff P. McDermott, Junjie Zhu, et al. Massively parallel digital transcriptional profiling of single cells. *Nature Communications*, 8:14049, 2017. doi: 10.1038/ncomms14049.
- Roland S. Zimmermann, Yash Sharma, Steffen Schneider, Matthias Bethge, and Wieland Brendel. Contrastive learning inverts the data generating process. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 12979–12990. PMLR, 2021.

Appendix A. Extraction, hierarchy definitions and reproducibility

The camera-ready artifact is `experiments/`. Its analysis reads immutable representation caches and writes a single JSON record with input SHA256 hashes, software versions, FastME options, scores and inferred empirical splits. The manuscript’s result macros, tables and three figures are generated from that record. Input data, reference definitions and extraction provenance are stored separately under `data/`. Tests cover known topologies, complementary splits, rerooting, row permutations, unresolved quartets, ties, random-label nulls, radial invariance, translation invariance of Euclidean distance, Newick parsing, and exact additive recovery by both estimators. No encoder is retrained or re-extracted for this audit. The extraction scripts document historical model identifiers; the saved caches, rather than a claim of bitwise reproducibility from changing upstream checkpoints, define this analysis.

Text. XLM-R-large (Conneau et al., 2020) embeds 1,012 FLORES-200 devtest sentences per language (NLLB Team, 2022) in float32, with a maximum token length of 256. Mean pooling excludes special tokens. The primary readout uses the cache label L12, which denotes transformer block 12 with zero-based indexing, or `hidden_states[13]`. Cached alternatives are L4, L8 and L18, plus pooling over the 80 most frequent nonspecial subwords, with mean-pool fallback for empty matches. Sentence frequencies use the full corpus. The language hierarchy consists of coded macrofamily, branch, genus and language paths; its pair distances are prefix-tree distances. URIEL family vectors are a separate reference used for the adjusted language analyses.

For language ℓ , the alignment divides representations by the RMS deviation of its training cloud, estimates an orthogonal matrix by five generalized Procrustes iterations using centered training clouds, then applies that matrix to the *uncentered*, scaled observations. The stored centroid is the test mean minus the grand mean over all aligned observations in the fitted population. Thus rotation fitting uses paired training sentences, while the origin includes test observations. The 36-language and 89-language alignments are separately fitted. Our additional centering subtracts the mean of the vectors in the evaluation population; it does not refit rotations. The source-deletion analysis holds these cached fits fixed, recomputes an evaluation mean when that readout calls for one, and refits the residual regression.

Speech. XLS-R-300m (Babu et al., 2022) embeds 100 utterances per language from FLEURS (Conneau et al., 2022). Time pooling excludes padding, then utterance vectors are averaged. Layers use `hidden_states[6,12,18,24]`, a convention distinct from the text cache. The primary speech row uses index 12; the final layer is evaluated explicitly rather than substituted for it. Reference paths group Romance, Germanic, Slavic and Indo-Aryan under Indo-European and retain the other coded families separately. Nearest-neighbor purity uses the family level.

Protein. The checkpoint is `facebook/esm2_t33_650M_UR50D` (Lin et al., 2023). Fifteen gene-name queries per vertebrate are sent to UniProt, preferring reviewed entries and otherwise taking the first returned sequence. Sequences are capped at 1,022 residues. Pooled embeddings for each gene are standardized coordinate-wise across species with available sequences; the 15 blocks are concatenated into a 19,200-dimensional species vector. Missing

blocks are zero. This retrieval does not verify one-to-one orthology. Missingness ranges from no blocks to 14 of 15, and can itself affect similarity. All species-specific counts and gene names are saved in the result record. The reference uses class/subgroup/species paths (e.g., Mammal/Euarch/human), not branch-length estimates or divergence dates. Its six nontrivial splits are a limited taxonomic target.

Vision. DINOv2-base (Oquab et al., 2023) supplies the final CLS representation after resizing images to 224×224 . All 100 test images per CIFAR-100 class are averaged (Krizhevsky, 2009). The 20 official superclasses are grouped into five authored metaclasses: animal, plant, house, outdoor and vehicle. The exact mapping and all leaf paths are in the JSON. The reference therefore has 25 scored semantic splits. We distinguish purity at these two grouping levels throughout.

Single-cell. Geneformer V2-104M is used with its transcriptome tokenizer and cell embedding extraction at `emb_layer=-1`; Theodoris et al. (2023) describes the Geneformer framework rather than this exact checkpoint version. The processed PBMC3k labels (Zheng et al., 2017) define eight cell-type centroids: four lymphoid types (CD4 T, CD8 T, B and NK), three myeloid types (CD14 monocytes, FCGR3A monocytes and dendritic cells), and megakaryocytes. The latter singleton supplies no nontrivial split. These are cell-type groupings, not lineage-tracing observations.

Appendix B. Complete cross-domain scores and random-label tests

Domain		r	Q_D	NN	Null NN	NJ splits	FM splits	Q_{NJ}	Q_{FM}
Text	E	0.232	0.566	0.483	0.051	8/36	10/36	0.624	0.631
Text	C	0.351	0.569	0.562	0.051	6/36	10/36	0.640	0.650
Speech	E	0.137	0.544	0.600	0.145	1/7	1/7	0.492	0.550
Speech	C	0.130	0.530	0.633	0.145	1/7	1/7	0.511	0.537
Protein	E	0.141	0.552	0.600	0.363	0/6	0/6	0.410	0.413
Protein	C	0.433	0.638	0.800	0.363	0/6	1/6	0.587	0.711
Vision	E	0.255	0.727	0.900	0.308	2/25	2/25	0.820	0.757
Vision	C	0.499	0.726	0.960	0.308	2/25	2/25	0.822	0.756
Single-cell	E	0.592	0.854	0.750	0.321	1/2	1/2	0.812	0.812
Single-cell	C	0.524	0.646	0.750	0.321	1/2	1/2	0.625	0.625

Table 2: E: Euclidean; C: cosine; FM: FastME. Q_D is direct distance quartet agreement, while Q_{NJ} and Q_{FM} score the reconstructed trees on the same reference-resolved quartets. Split columns give recovered/available counts. NN uses text branch, speech family, protein class, vision metaclass and cell grouping.

The quartet samples are uniform without replacement (seed 0). Reference-unresolved quartets are excluded after sampling. Ties use tolerance $10^{-10} \max(1, \max D)$; a prediction tied among k pairings receives $1/k$ credit if the reference pairing is among them. Nearest-neighbor ties are averaged similarly. The random-label test permutes the rows and columns of the representation matrix together, holding reference paths and group sizes fixed. The number of resolved quartets is not an effective sample size for inference.

Domain		Resolved	Examined	p_r	p_Q	p_{NN}	Mean residual	Max residual
Text	E	24700	40000	0.0005	0.0005	0.0005	0.0092	0.0655
Text	C	24700	40000	0.0005	0.0005	0.0005	0.0205	0.1497
Speech	E	19557	27405	0.1144	0.0005	0.0005	0.0308	0.1517
Speech	C	19557	27405	0.1394	0.0005	0.0005	0.0439	0.2343
Protein	E	10337	12650	0.2514	0.0005	0.0210	0.0146	0.0694
Protein	C	10337	12650	0.0005	0.0005	0.0005	0.0373	0.1926
Vision	E	26568	40000	0.0005	0.0005	0.0005	0.0111	0.0846
Vision	C	26568	40000	0.0005	0.0005	0.0005	0.0247	0.1494
Single-cell	E	48	70	0.0035	0.0035	0.0340	0.0196	0.0633
Single-cell	C	48	70	0.0035	0.0530	0.0340	0.0248	0.0830

Table 3: Source-label permutation results. The minimum attainable reported p is 1/2001. For each tested quartet, the four-point residual is $(s_{(3)} - s_{(2)})/(2 \max D)$, where s are the three pair sums. “Max” is a sampled maximum for text/vision and exhaustive for the smaller populations; the mean is not Gromov’s maximum four-point constant. For cosine, these remain descriptive dissimilarity residuals. No multiple-comparison-adjusted discovery claim is made.

Appendix C. Full readout sweeps

Layer		Raw	Raw, centered	Align 89	Align 89, centered	Align 36	Align 36, centered
4	E	0.238	0.238	0.133	0.133	0.117	0.117
4	C	0.169	0.449	0.321	0.378	0.358	0.358
8	E	0.172	0.172	0.072	0.072	0.068	0.068
8	C	0.100	0.402	0.272	0.355	0.365	0.365
12	E	0.084	0.084	0.090	0.090	0.078	0.078
12	C	0.053	0.432	0.306	0.442	0.431	0.436
18	E	0.126	0.126	0.005	0.005	-0.004	-0.004
18	C	0.077	0.399	0.181	0.261	0.256	0.255

Table 4: Adjusted language correlation on the same 36-language population. “Centered” means subtracting the evaluation-centroid mean; aligned variants otherwise use their stored origin. E: Euclidean; C: cosine. The target is URIEL family distance, with imputed syntax as a descriptive control.

For example, the source-deletion range of the centered raw L12 cosine association in the 36-language population is [0.381, 0.468]. This records dependence on individual sources while leaving the fitted encoder and cached observations fixed. It does not quantify extraction uncertainty or replace an independent evaluation population.

Appendix D. Additional language analyses

Within-family association. The six-family analysis covers 36 languages and 94 within-family pairs. It averages Euclidean distances from four text layers and two poolings, each

Layer		Raw	Raw, centered	Align 89	Align 89, centered
4	E	0.228	0.228	0.067	0.067
4	C	0.216	0.161	0.115	0.116
8	E	0.210	0.210	0.110	0.110
8	C	0.191	0.137	0.107	0.109
12	E	0.216	0.216	0.152	0.152
12	C	0.203	0.171	0.123	0.123
18	E	0.207	0.207	0.155	0.155
18	C	0.192	0.152	0.095	0.095

Table 5: The corresponding 89-language evaluation. These associations describe a different population; they should not be compared with the 36-language results as a pure alignment effect.

Layer	PR	Mean cosine	Q_D	NN	NJ splits	FM splits	Q_{NJ}	Q_{FM}
6	10.32	0.9676	0.432	0.500	1/7	1/7	0.481	0.448
12	9.55	0.9827	0.530	0.633	1/7	1/7	0.511	0.537
18	11.58	0.9766	0.513	0.600	1/7	1/7	0.493	0.471
24	2.31	0.9979	0.357	0.133	0/7	0/7	0.355	0.359

Table 6: Speech layer sweep using cosine dissimilarity. PR and mean pairwise cosine are computed on the cached source centroids. Euclidean results and all source-deletion estimates for text are included in the machine-readable record.

divided by its mean within-family distance, using separately fitted family frames. Syntax distances use either dense imputed features or the features co-observed for a pair; the latter uses mean absolute differences over at least eight jointly observed features. Thus the two controls also differ in their distance construction, not just missingness. Restricting both to the same pair mask removes the large apparent effect attributed to imputation in an unmatched comparison. Partial correlations residualize both variables on the control and an intercept.

Control and pair set	Pairs	Partial r	Source deletion range
Imputed, all pairs	94	0.264	[0.192, 0.375]
Imputed, matched pairs	81	0.352	[0.271, 0.459]
Observed, matched pairs	81	0.349	[0.249, 0.440]
WALS only	57	0.330	[0.246, 0.477]
Reverse partial, observed	81	-0.026	[-0.112, 0.096]

Table 7: Matched and unmatched controls. All n values here count dependent pairs, not independent languages. The reverse partial is geometry versus observed typology given genealogy. Source-deletion ranges are descriptive.

Divergence-time association. We retain a secondary analysis on 23 languages from Romance, Slavic, Germanic and Turkic, producing 58 within-family pairs. Pairwise geometry and the hand-specified time axis are standardized separately within each family before pooling. Three date sets are stored in the artifact: the historical analysis’s original values, a corrected set using published anchors, and an alternative anchor sensitivity (Chang et al., 2015; Holman et al., 2011; Savelyev and Robbeets, 2020). These are coarse node assignments, not independently measured pairwise dates. Lexically informed dates can share evidence with language similarity, so the association is not an independent calibration of a historical clock.

Dates	r	ρ	p	Deletion range	r w/o Germ.	p w/o Germ.
Original	0.514	0.355	0.0005	[0.429, 0.589]	0.729	0.0005
Cited	0.492	0.356	0.0010	[0.406, 0.564]	0.674	0.0005
Alternative	0.467	0.346	0.0010	[0.380, 0.543]	0.674	0.0005

Table 8: Date sensitivity with Pearson r , pooled Spearman ρ , and within-family source-label permutation p (two-sided, 2,000 draws). Removing Germanic is a reported sensitivity analysis, not selection of a preferred population. Both geometry and dates are standardized within family.

Under the corrected dates, within-family Pearson correlations are 0.532 (Romance), 0.664 (Slavic), 0.171 (Germanic), and 0.945 (Turkic). Permutation shuffles source labels within each family and reconstructs its time-distance vector; individual pairs are never shuffled independently.

Other multilingual encoders. The cached mBERT, XLM-R-base and LaBSE (Feng et al., 2022) analyses use two layers (cache labels L6 and L9), family-specific alignment and the same within-family pair definitions. Table 9 reports descriptive adjusted correlations and corrected-date clock associations. Differences among these models do not isolate capacity: architecture, pretraining and pooling differ, and the XLM-R-large ensemble uses more configurations.

Encoder	Imputed 94	Imputed 81	Observed 81	Clock r	Clock p
mbert	0.097	0.222	0.212	0.479	0.0005
xlmr_base	0.126	0.205	0.220	0.363	0.0020
labse	0.212	0.256	0.258	0.352	0.0055

Table 9: Additional encoders. The three adjusted-correlation columns distinguish pair availability and control construction. Clock p uses within-family source permutations, not pair permutations.

Cross-script proximity. We report all four focal pairs, including readout-sensitive cases, rather than retain only stable ones. Percentiles rank distances among all 3,916 language pairs; lower means closer.

Pair	Single E	Single C	Ensemble E	Surface
Hindi–Urdu	4.400	1.100	14.683	97.300
Bengali–Hindi	45.700	21.600	51.226	97.700
Kazakh–Turkish	19.000	34.400	3.550	76.500
Arabic–Hebrew	39.300	40.000	25.919	64.900

Table 10: The single configuration is global mean-pool L12; the ensemble averages eight globally aligned, mean-distance-normalized Euclidean matrices. The surface baseline is character 3–5-gram TF–IDF cosine dissimilarity. Percentiles are descriptive, without a cross-script invariance claim.

The surface baseline has median distance 0.986; a fraction 0.314 of pairs exceed 0.999. Near this ceiling, large percentile differences can reflect very small distance differences. Hindi–Urdu is close under the single geometric readouts, but neither that example nor the saturated surface ranking establishes general script invariance.

Appendix E. Synthetic sweep and the scope of supplementary controls

k_N	k_C		r	Q_D	S_{NJ}	S_{FM}	Q_{NJ}	Q_{FM}	Deep S_{NJ}
0	0	E	0.762	0.808	0.961	0.969	0.977	0.977	0.960
0	0	C	0.808	0.797	0.948	0.964	0.951	0.959	0.880
0.5	0	E	0.184	0.703	0.333	0.425	0.930	0.948	0.760
0.5	0	C	0.808	0.797	0.948	0.964	0.951	0.959	0.880
1	0	E	0.046	0.533	0.123	0.160	0.802	0.846	0.120
1	0	C	0.808	0.797	0.948	0.964	0.951	0.959	0.880
1.5	0	E	0.014	0.450	0.088	0.109	0.612	0.591	0.000
1.5	0	C	0.808	0.797	0.948	0.964	0.951	0.959	0.880
2	0	E	0.005	0.411	0.070	0.088	0.482	0.533	0.000
2	0	C	0.808	0.797	0.948	0.964	0.951	0.959	0.880
0	0.1	E	0.721	0.777	0.965	0.971	0.989	0.979	1.000
0	0.1	C	0.764	0.768	0.946	0.965	0.954	0.967	0.920
0	0.2	E	0.643	0.714	0.942	0.956	0.915	0.879	0.800
0	0.2	C	0.676	0.705	0.916	0.941	0.874	0.887	0.720
0	0.35	E	0.429	0.572	0.518	0.626	0.717	0.632	0.000
0	0.35	C	0.449	0.568	0.557	0.628	0.697	0.629	0.000
0	0.5	E	0.211	0.450	0.033	0.037	0.418	0.425	0.000
0	0.5	C	0.229	0.450	0.037	0.036	0.428	0.420	0.000

Table 11: Five-seed means for every declared perturbation setting. S scores all 253 unique internal splits; “Deep” scores the five unique splits from the first two levels. E: Euclidean; C: cosine. Individual seed scores, full-tree exactness and normalized RF distances are saved in the JSON. Seed variability appears in Figure 3, not as an uncertainty estimate for the empirical populations.

Seeds are 0–4. Independent random streams for donors (seed plus 1,000) and radii (seed plus 2,000) make perturbations identical across doses within a seed. This explicitly specified coupling differs from historical sweeps that reused a generator’s evolving random state. The implemented order is mixing followed by radial scaling; the reported sweep varies one at a time, so no combined-cell or historical causal-order claim is made. The Brownian generator and the exact-distance oracle are separate controls. Historical deep-clade scores counted six rooted groups, including complementary root groups twice; they are not interchangeable with the unique all-split scores reported here.

E.1. Archived cloud statistics

Archived statistic	Layer	Partial r
paired displacement magnitude	8	0.223
paired displacement magnitude	12	-0.196
PCA-40 Bures covariance	8	-0.017
PCA-40 Bures covariance	10	-0.277
PCA-40 Bures covariance	12	-0.265
PCA-40 Bures covariance	16	-0.276

Table 12: Quantified archived language cloud analyses, residualizing on centroid distance. The displacement-magnitude values are available to three decimals in the archived displacement-magnitude summary; covariance values are recomputed from the archived covariance pair-level CSV. These are association tests on 94 within-family pairs, not additional tree-reconstruction methods.

The first analysis summarizes variation in aligned paired sentence displacements beyond the centroid offset. The second computes Bures distances between covariances of sentence-meaning-removed form residuals after a joint PCA reduction to 40 dimensions. Their mixed signs and layer dependence do not establish that higher moments are uninformative. Neither analysis tests arbitrary multimodal distributions, learned subspaces or nonlinear probes. We therefore report them as limited supplementary evidence and make no general exclusion claim about distributional geometry. Training-objective sweeps without comparable archived quantitative evaluations are not used to support a conclusion.