Beyond cognacy

Gerhard Jäger
University of Tübingen, Germany
gerhard.jaeger@uni-tuebingen.de

Abstract

Computational phylogenetics has become an established tool in historical linguistics, with many language families now analyzed using likelihood-based inference. However, standard approaches rely on expert-annotated cognate sets, which are sparse, labor-intensive to produce, and limited to individual language families. This paper explores alternatives by comparing the established method to two fully automated methods that extract phylogenetic signal directly from lexical data. One uses automatic cognate clustering with unigram/concept features; the other applies multiple sequence alignment (MSA) derived from a pair-hidden Markov model. Both are evaluated against expert classifications from Glottolog and typological data from Grambank. Also, the intrinsic strengths of the phylogenetic signal in the characters are compared. Results show that MSA-based inference yields trees more consistent with linguistic classifications, better predicts typological variation, and provides a clearer phylogenetic signal, suggesting it as a promising, scalable alternative to traditional cognate-based methods. This opens new avenues for global-scale language phylogenies beyond expert annotation bottlenecks.

1 Introduction

Originally developed in computational biology, quantitative methods for phylogenetic reconstruction using likelihood-based inference frameworks have now gained widespread acceptance in comparative linguistics. This is evident from the growing number of computational phylogenies proposed for some of the world's largest language families, including Dravidian (Kolipakam et al., 2018), Sino-Tibetan (Sagart et al., 2019), and Indo-European (Heggarty et al., 2023). Moreover, fully automated approaches — where even cognate identification is performed algorithmically — have demonstrated a surprising degree of robustness (Rama et al., 2018). In contrast to the pre-computational era of historical linguistics, where such detailed reconstructions were rare, the generation of fully resolved phylogenies with branch lengths and, in some cases, estimated divergence dates has now become a standard practice in studies of language evolution.

Despite the increasing recognition of computational language phylogenies as a useful addition to the comparative linguistics toolkit, skepticism remains prevalent. A key concern raised by critics is that phylogenetic analyses are often based on cognate sets—groups of historically related words—extracted from semantically aligned word lists. Since these cognate sets are based on expert annotations, they are sparse, labor-intensive to acquire, and raise concerns regarding replicability.

Another limitation of phylogenetic inference based on cognate classes is that it is by definition constrained to individual language families. There is legitimate interest in automatically inferred trees spanning larger collections of languages, perhaps from the entire world. Such trees provide information about the strength of evidence for putative macro-families (Jäger, 2015; Akavarapu and Bhattacharya, 2024). Furthermore, they are useful for downstream tasks such as the statistical modeling of global language evolution (Bentz et al., 2018; Bouckaert et al., 2022).

The literature contains several proposed workflows for extracting character matrices from word lists without cognate annotations, which can then used as input for likelihood-based phylogenetic inference. This paper presents a comparison of cognate-based phylogenetic inference with two such proposals, the one by Jäger (2018) and the one by Akavarapu and Bhattacharya (2024). These methods are evaluated in three ways: (1) by comparing the inferred phylogenies with the Glottolog expert classification (Hammarström et al., 2024),

Proceedings of the 7th Workshop on Research in Computational Linguistic Typology and Multilingual NLP (SIGTYP 2025), pages 52–60

(2) how well the inferred phylogenies fit to the typological features from (Skirgård et al., 2023), and (3) an estimation of the strength of the phylogenetic signal in the data, which is inferred with the software *PyPythia* (Haag et al., 2022).

2 Materials and Methods

2.1 Materials

Word lists were obtained from Lexibank[1]; List et al. 2023). These datasets contain lexical entries, including the language they belong to, their meaning, form in IPA transcription, and often a manual cognate annotation. The datasets are curated by the Lexibank community and are available in a standardized format, which makes them suitable for computational analyses.

In a first step, 135 Lexibank dataset were selected. In total, this amounts to 2,486,845 lexical entries from 6,845 languages (identified by glottocodes).

For the purpose of evaluation, typological features were obtained from Grambank[2]. This results in 355,097 binary entries from 2,467 languages and 195 typological features.

A subset of Lexibank data was selected according to the following criteria:

- The entry comes from a language with a Glottocode that is present in the Grambank data.
- The entry has an entry for its meaning (`Concepticon_Gloss`) and a manual cognate annotation (`Cognateset_ID`).
- The meaning comes from the 110 concepts with the largest coverage.

This leaves 113,671 entries from 928 languages. For further processing, the IPA transcriptions were converted to the ASJP alphabet using the python package *lingpy* (List and Forkel, 2024).

Constraining the Grambank data to these 928 languages leaves 138,878 binary data points from all 195 features.

The gold standard tree was obtained from Glottolog.[3]

[1] https://github.com/lexibank

[2] https://github.com/grambank/grambank

[3] https://zenodo.org/records/10804582/files/glottolog/glottolog-cldf-v5.0.zip

2.2 Methods

The overall workflow consists of the following steps:

1. Phylogenetic inference
   (a) Generate a binary character matrix from the Lexibank data.
   (b) Infer a phylogenetic tree from this character matrix.
2. Evaluation
   (a) Compare the inferred phylogenetic tree with the Glottolog expert classification.
   (b) Compare the inferred phylogenetic tree with the Grambank typological features.
   (c) Assess the strength of the phylogenetic signal in the data.

Three different methods were used to generate a binary character matrix: (1) binarized expert-annotated cognate classes, (2) a combination of automatic cognate clustering and unigram/concept features as described in Jäger (2018), and (3) a variant of the method developed by Akavarapu and Bhattacharya (2024) using multiple sequence alignment.

2.2.1 Expert-annotated cognate classes (cc)

Here we use the method introduced by Ringe et al. (2002) and Gray and Atkinson (2003). Each cognate class is treated as a character. A language is coded as 1 if it has a cognate in the class, 0 if it has a different cognate class for the same concept, and missing if it has no cognate for the concept. This results in a matrix with 928 rows and 25,913 columns.

Since each cognate class is, by definition, confined to a single language family, this character matrix contains no signal beyond the family level.

In the tables and figures below, this method is referred to as cc (for cognate classes).

2.2.2 Automatic cognate clustering and unigram/concept features (PMI)

The workflow proposed by Jäger (2018) was replicated. This approach uses two types of characters.

- Binarized cognate classes obtained via automatic cognate clustering. This involves (1) supervised training of a Support Vector Machine classifier which takes a pair of words

and predicts the labels 1 (cognate) or 0 (non-cognate), using manual cognate classification for supervision, (2) creating a distance matrix for all entries for a given concept from the 100 concepts defined above, and (3) clustering the distance matrix using the *label propagation* algorithm (Raghavan et al., 2007).

- Unigram/concept characters. For each combination of a concept c and an ASJP sound class s, a language is coded as 1 if it has a word for concept c that contains sound class s, missing if it has no word for concept c, and 0 otherwise.

This resulted in a matrix with 928 rows and 41,013 columns.

Since the *pointwise mutual information* between sound classes plays an essential role in this workflow, the method is referred to as PMI.

2.2.3 Multiple sequence alignment (MSA)

The method by Akavarapu and Bhattacharya (2024) was used as starting point, but the present approach differs in various aspects. The method is based on the following steps:

In a first step, pairwise distances between languages in the full lexibank dataset were computed using the Levenshtein distance on the ASJP transcriptions and aggregating according to the method described in (Jäger, 2018). Language pairs with a distance below 0.7 were considered as *probably related*, using the same heuristics as Jäger (2018). There are 172,681 such language pairs. All word pairs from such a language pair sharing their meaning are treated as *potential cognates*. There are 90,565,486 such word pairs. An equal number of random word pairs were sampled as probable non-cognates. Potential cognates were assigned the label 1 and probable non-cognate the label 0.

In a second step, a classifier was trained on the potential cognates and non-cognates. The classifier consists of a *pair-Hidden Markov Model* (pHMM) (Durbin et al., 1998) and a logistic-regression layer. The classifier was trained for one epoch using the Adam optimizer. The resulting parameters of the pHMM were used in the next step.

A pHMM defines a probability distribution over pairs of aligned sequences of sound classes. This involves (1) emission probabilities for all pairs of sound classes that are matched in the alignment, (2) emission probabilities for individual sound classes if they are aligned with a gap, and (3) transition probabilities between the hidden states *match*, *gap in string 1*, *gap in string 2*, and *final state*.

It is instructive to inspect the emission probabilities in the trained model. In Table 1 the ten sound classes with the highest probability of being matched with /p/ are shown for illustration, together with their log-probabilities. This ranking is in good agreement with linguistic intuitions about potential sound correspondences.

Sound class	Log-probability
p	−2.39
f	−16.35
b	−18.85
v	−23.26
h	−24.03
L	−25.11
g	−27.67
7	−29.74
C	−29.95
I	−30.78

Table 1: The ten sound classes with the highest probability of being matched with /p/ in the trained pHMM, along with their log-probabilities.

Sound class	Log-probability
c	-0.86
j	-1.17
L	-1.54
1	-2.94
I	-8.06
h	-9.37
7	-9.60
i	-10.14
y	-10.24
T	-10.33

Table 2: The ten sound classes with the highest probability of being matched with a gap in the trained pHMM, along with their log-probabilities.

A high probability here is to be interpreted as a high likelihood that instances of these sound classes participate either in insertion or deletion.

The trained pHMM assigns a probability to each pair of aligned sequences. Via the forward algorithm, the probability of a pair of sequences is computed as the sum of the probabilities of all possible alignments between these sequences.

Following Durbin et al. (1998), a null-model

was trained additionally that assigns individual probabilities to both sequences, disregarding the order of sound classes. The log-odds ratio of a pair of words of being generated by the pHMM vs. the null model can interpreted as a measure of the similarity of the two words.

To illustrate this, a collection of ten words were chosen at random from the dataset which all have an edit distance of 1 to the word *baba*, and their log-odds ratios with respect to *baba* were computed. The results are shown in Table 3.

word	log-odds
babae	98.26
babau	96.31
bIba	95.73
bawa	87.55
zaba	85.51
raba	74.58
maba	73.52
eaba	73.50
xaba	71.78
naba	70.94

Table 3: Ten randomly chosen words with an edit distance of 1 from *baba*, alongside with the predicted log-odds to *baba*.

This ranking illustrates that the log-odds predicted by the trained pHMM are consistent with linguistic intuitions about potential cognacy.

In a third step, the trained pHMM was used in combination with the Viterbi algorithm to obtain pairwise sequence alignments for all synonymous word pairs from different languages within the smaller dataset of 928 languages and 110 concepts.

In a fourth step, the pairwise sequence alignments were aggregated to a *multiple sequence alignment* (MSA) using the *T-Coffee* algorithm (Notredame et al., 2000).

Note that all reflexes of a given concept are aligned within a single MSA, regardless of cognacy. Such an MSA implicitly contains information both about cognacy and about sound correspondences.

An example (for a much smaller dataset) is shown in Table 4 for illustration. These are the reflexes of the concept *louse* from the Tungusic languages in the dataset.[4]

As can be seen from this example, the MSA contains information about cognacy, but also about sound correspondences. For example, a `t` in the first column is a proxy for the cognate class `16_lousen-38`. The sound classes `k` and `q`, on the other hand, both correspond to the cognate class `16_lousen-37`, and they additionally reflect a sound change. In column 4, however, the cognate class `16_lousen-38` is split into two sound classes, `k` and `q`, reflecting a sound change. The presence of a sound class, as opposed to a gap, is a proxy of that cognate class. Put differently, binary characters corresponding to a gap are flipped by switching 0s and 1s.

In a fifth step, the MSA was converted to a binary matrix. Two binarization methods were used simultaneously. For a given column in an MSA, a character was created for the presence of a sound class. For column 4 in Table 4, e.g., this character has value 1 for Nanai, Orok and Ulch, and 0 for the other languages. Additionally, for each sound class type in a column, a different character was created. In the example, there are two such characters, one for `k` and one for `q`. The first has value 1 for Nanai and Orok and 0 otherwise, while the second has value 1 for Ulch and 0 otherwise. Languages for which the data do not contain a reflex for a given concept are coded as missing for all relevant characters. If a language has multiple reflexes for a given concept, the maximum value is chosen.

Applying this workflow to all concepts and concatenating the resulting matrices yields the final character matrix 928 rows and 46,409 columns.

As mentioned above, this workflow builds on the method by Akavarapu and Bhattacharya (2024), but differs in various aspects. The mentioned work (1) uses Dolgopolsky sound classes instead of ASJP, (2) finds the MSA using CLUSTALW2 (Larkin et al., 2007) instead of T-Coffee, and (3) omits the binarization steps, working with a multistate model of evolution for phylogenetic inference.

In the tables and figures this method is referred to as MSA.

2.2.4 Phylogenetic inference

We performed phylogenetic inference using *raxml-ng* (Kozlov et al., 2019), which implements maximum-likelihood estimation. The GTR+G model (generalized time-reversible model with

[4]The data are taken from `https://zenodo.org/records/13163376`, which is based on (Oskolskaya et al., 2021).

Language	Cognateset_ID	1	2	3	4	5	6	7	8
Even	16_lousen-37	k	-	u	-	m	-	k	e
Kilen	16_lousen-37	q	h	u	-	m	I	k	I
Negidal	16_lousen-37	k	-	u	-	m	-	k	I
Oroch	16_lousen-37	k	-	u	-	m	-	-	I
Udihe	16_lousen-37	k	-	u	-	m	u	x	I
Nanai	16_lousen-38	t	-	i	k	t	-	-	I
Orok	16_lousen-38	t	-	i	k	t	-	-	I
Ulch	16_lousen-38	t	-	i	q	t	-	-	I

Language	sound class	k	q
Even	0	0	0
Kilen	0	0	0
Negidal	0	0	0
Oroch	0	0	0
Udihe	0	0	0
Nanai	1	1	0
Orok	1	1	0
Ulch	1	0	1

Table 4: Example of a multiple sequence alignment. Alignment cells are shaded to indicate different cognate sets. (left) Binarized version of column 4. (right)

gamma-distributed rates) was used for all analyses. This means that gain rates and loss rates can be different, and that the mutation rates of the different characters can differ but are drawn from the same gamma distribution. The parameters of this distribution are estimated from the data.

Using the standard settings of *raxml-ng*, 20 maximum likelihood tree searches were performed, ten of them starting from random trees and ten from maximum-parsimony trees. The tree with the highest likelihood was chosen as the final result.

2.2.5 Evaluation

Evaluation was conducted on three types of datasets:

- the full dataset of 928 languages,
- 100 samples of 100 languages each, which are drawn at random without replacement from the full dataset, and
- a collection of 14 language families, each with at least 10 languages.

For each of these groups of datasets, the following evaluations were performed:

Comparison with Glottolog The Glottolog classification of the languages in a dataset can be represented as a phylogenetic tree with polytomies, i.e., with nodes containing more than two daughters. This Glottolog tree serves as gold standard. To assess the degree of agreement between the gold standard and the inferred phylogenies, the *generalized quartet distance* (GQD) was deployed, as first proposed by Pompei et al. (2011). This distance is defined as the fraction of quartets (i.e., sets of four languages) that are (a) resolved in both trees, and (b) resolved differently in the two trees. The GQD ranges from 0 (perfect agreement) to 0.67 (chance level). The GQD was computed using the software *QDist*, which can be obtained from https://birc.au.dk/software/qdist/.

Fit with Grambank The hypothesis is assumed that the values of the Grambank features evolve along a phylogeny in the same way as the lexical characters described earlier in this section. The degree of fit of the inferred phylogenies with the Grambank features was assessed by (1) using the inferred phylogeny and estimating the branch lengths, mutation rates and rate heterogeneity via Maximum Likelihood, and (2) computing the *Akaike Information Criterion* (AIC). A lower AIC value indicates a better fit.

ML inference and AIC computation were also performed with *raxml-ng*.

For the groups of random samples and of language families, AIC values were normalized to mean 0 to facilitate comparison.

Phylogenetic difficulty The strength of the phylogenetic signal in the data was assessed using the software *PyPythia* (Haag et al., 2022). The authors define a measure of signal strength that uses 100 maximum likelihood tree searches and quantifies the degree of agreement between the inferred trees. The software *PyPythia* implements a machine learning algorithm that predicts this difficulty from various properties of the character matrix, such as entropy and sites-over-taxa ratio, and maximum-parsimony tree inference, with high precision and comparatively low computational cost. The measure ranges from 0 (little difficulty, i.e., strongest signal) to 1 (very difficult, i.e., no signal).

Method	GQD (Glottolog)	AIC (Grambank)	difficulty
Cognate classes	0.188	105.340	0.59
PMI	0.062	104.903	0.63
MSA	**0.042**	**104,752**	**0.45**

Table 5: Evaluation of the full dataset. GQD = Generalized Quartet Distance (lower is better; ranges from 0 for perfect fit to 0.67 for chance level); AIC = Akaike Information Criterion for typological model fit (lower is better; absolute values are not interpretable in isolation but differences are meaningful); difficulty = phylogenetic difficulty estimated by PyPythia (lower is better; ranges from 0 for strong phylogenetic signal to 1 for absent signal).

Method	μ GQD	σ GQD	μ AIC	σ AIC	μ difficulty	σ difficulty
Cognate classes	0.227	0.077	151	115	0.486	0.030
PMI	0.095	0.030	−28	69	0.326	0.032
MSA	**0.048**	0.015	**−123**	66	**0.294**	0.021

Table 6: Evaluation of the 100 random samples (μ: sample mean; σ: sample standard deviation).

Method	μ GQD	σ GQD	μ AIC	σ AIC	μ difficulty	σ difficulty
Cognate classes	0.223	0.130	**−1.73**	17.01	0.401	0.164
PMI	0.221	0.109	3.42	20.43	0.280	0.187
MSA	**0.218**	0.109	−1.69	14.30	**0.203**	0.159

Table 7: Evaluation of the 14 largest language families (μ: sample mean; σ: sample standard deviation).

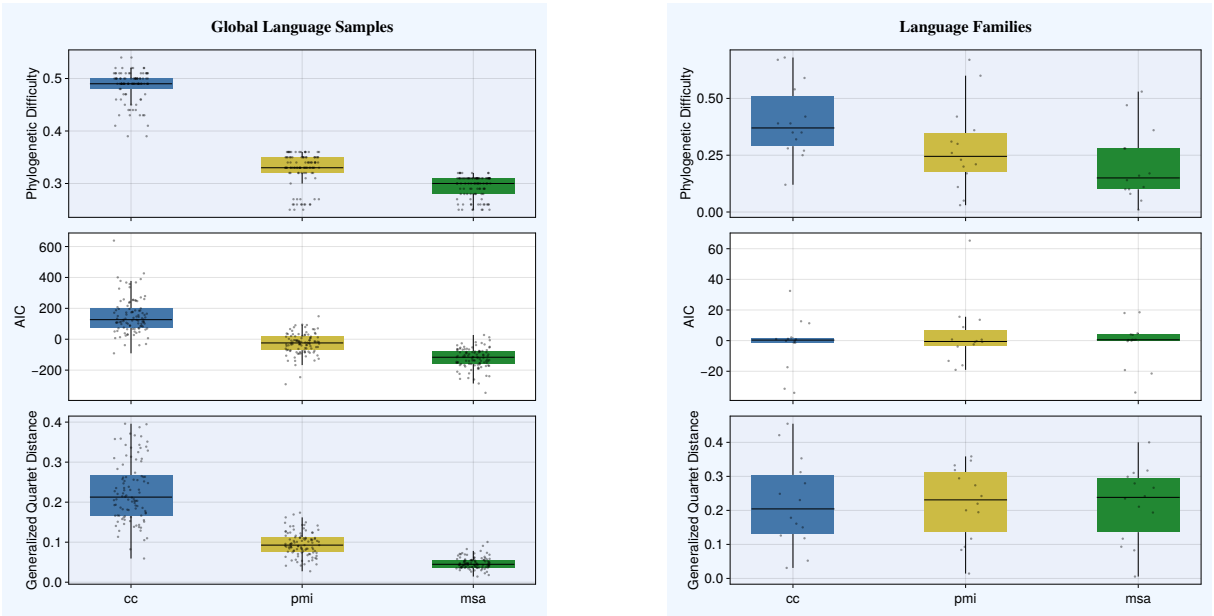

Figure 1: **Left panel:** Comparison of methods across three evaluation metrics for the 100 random samples. The boxplots show distribution per method, while the overlaid points represent individual samples. **Right panel:** Comparison of methods across three evaluation metrics for the 14 largest language families. The boxplots show distribution per method, while the overlaid points represent individual samples.

3 Results

The evaluation results for the entire dataset are shown in Table 5. Table 6 shows the aggregated results for the 100 random samples. They are visualized in the left panel of Figure 1.

The aggregated evaluation results for the 14 largest language families are shown in Table 7. They are visualized in the right panel of Figure 1.

The results for the individual families are given in Table 8.

When focusing on phylogenetic inference at the level of individual families, we find a considerable variation between families. This applies both to the numerical evaluation results and the relative ranking of the three methods considered here. The MSA method tends to produce the lowest phylogenetic difficulty, while there is no discernible trend regarding the fit to Glottolog and to Grambank.

This picture changes considerably when we focus on datasets covering languages from many different families. Here, the MSA method consistently outperforms the other two methods. This is particularly evident in the comparison with Glottolog, where the MSA method yields the lowest GQD values. The MSA method also leads to the lowest AIC values, indicating a better fit to the Grambank typological features. The phylogenetic difficulty is also lowest for the MSA method.

4 Discussion

These findings suggest that the MSA method is a promising alternative to traditional cognate-based methods. It is competitive with the more labor-intensive method based on manual cognate annotations, as well as the method using automatically detected cognate classifications, when considering individual language families. For global datasets, the MSA method clearly outperforms the other two methods. This is particularly evident in the comparison with Glottolog, where the MSA method yields the lowest GQD values. The MSA method also tends to produce the lowest AIC values, indicating a better fit to the Grambank typological features. The phylogenetic difficulty is also lowest for the MSA method.

Limitations

The two evaluation methods that quantify the fit of the inferred trees to empirical data only assess the quality of the inferred tree **topologies**. Future work will need to address the question how well the inferred branch lengths and divergence dates correspond to the true values. This is a challenging task, as the true values are unknown. It is expected that the usefulness for downstream tasks is a suitable proxy.

Data and Code Availability

The code used in this study is available at `https://codeberg.org/profgerhard/sigtyp2025_code/`.

Acknowledgments

This research was supported by the DFG Centre for Advanced Studies in the Humanities Words, Bones, Genes, Tools (DFG-KFG 2237) and by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (Grant agreement 834050).

References

V.S.D.S.Mahesh Akavarapu and Arnab Bhattacharya. 2024. A likelihood ratio test of genetic relationship among languages. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2559–2570, Mexico City, Mexico. Association for Computational Linguistics.

Christian Bentz, Dan Dediu, Annemarie Verkerk, and Gerhard Jäger. 2018. The evolution of language families is shaped by the environment beyond neutral drift. *Nature Human Behaviour*, 2(11):816–821.

Remco Bouckaert, David Redding, Oliver Sheehan, Thanos Kyritsis, Russel Gray, Kate E Jones, and Quentin Atkinson. 2022. Global language diversification is linked to socio-ecology and threat status. *SocArXiv*.

Richard Durbin, Sean R Eddy, Anders Krogh, and Graeme Mitchison. 1998. *Biological sequence analysis: probabilistic models of proteins and nucleic acids*. Cambridge university press.

Russell D Gray and Quentin D Atkinson. 2003. Language-tree divergence times support the anatolian theory of indo-european origin. *Nature*, 426(6965):435–439.

Julia Haag, Dimitri Höhler, Ben Bettisworth, and Alexandros Stamatakis. 2022. From easy to hopeless—predicting the difficulty of phylogenetic analyses. *Molecular biology and evolution*, 39(12):msac254.

Harald Hammarström, Robert Forkel, Martin Haspelmath, and Sebastian Bank. 2024. Glottolog 5.1. *Leipzig: Max Planck Institute for Evolutionary Anthropology.(Available online atglottolog. org, Accessed on2024-11-29.)*, 10.

Paul Heggarty, Cormac Anderson, Matthew Scarborough, Benedict King, Remco Bouckaert, Lechosław Jocz, Martin Joachim Kümmel, Thomas Jügel, Britta Irslinger, Roland Pooth, Henrik Liljegren, Richard F.

Family	Metric	cc	pmi	msa
Afro-Asiatic	GQD	0.455	**0.195**	0.211
	PhyDiff	0.420	0.260	**0.100**
	AIC	12.635	8.841	**-21.476**
Arawakan	GQD	**0.280**	0.346	0.317
	PhyDiff	0.250	0.110	**0.080**
	AIC	0.868	**-0.845**	-0.023
Atlantic-Congo	GQD	**0.150**	0.220	0.235
	PhyDiff	0.680	0.670	**0.530**
	AIC	**-34.076**	15.572	18.504
Austroasiatic	GQD	**0.052**	0.093	0.093
	PhyDiff	**0.280**	0.310	0.280
	AIC	11.338	**-16.115**	4.777
Austronesian	GQD	**0.161**	0.200	0.194
	PhyDiff	0.670	0.600	**0.470**
	AIC	-31.419	65.299	**-33.880**
Chibchan	GQD	**0.118**	0.332	0.310
	PhyDiff	0.350	0.360	**0.110**
	AIC	-0.168	**-3.895**	4.063
Dravidian	GQD	0.312	0.242	**0.242**
	PhyDiff	0.390	0.170	**0.100**
	AIC	1.196	**-0.862**	-0.334
Indo-European	GQD	0.031	0.014	**0.005**
	PhyDiff	0.320	0.210	**0.140**
	AIC	1.065	**-19.079**	18.015
Pama-Nyungan	GQD	**0.178**	0.359	0.299
	PhyDiff	0.540	0.420	**0.360**
	AIC	**-17.375**	13.601	3.774
Sino-Tibetan	GQD	**0.230**	0.318	0.279
	PhyDiff	0.590	0.300	**0.280**
	AIC	32.455	-13.239	**-19.216**
Tucanoan	GQD	0.421	**0.274**	0.400
	PhyDiff	0.270	0.030	**0.010**
	AIC	**-1.364**	0.758	0.607
Tupian	GQD	0.353	0.294	**0.266**
	PhyDiff	0.390	0.200	**0.160**
	AIC	-0.187	**-0.280**	0.467
Turkic	GQD	0.249	**0.117**	**0.117**
	PhyDiff	0.350	0.230	**0.170**
	AIC	**-1.266**	0.647	0.618
Uto-Aztecan	GQD	0.126	0.084	**0.083**
	PhyDiff	0.120	**0.050**	0.050
	AIC	2.098	**-2.485**	0.388

Table 8: Evaluation of the 14 largest language families. The best value for each family is highlighted in bold.

Strand, Geoffrey Haig, Martin Macák, Ronald I. Kim, Erik Anonby, Tijmen Pronk, Oleg Belyaev, Tonya Kim Dewey-Findell, and 14 others. 2023. Language trees with sampled ancestors support a hybrid model for the origin of Indo-European languages. *Science*, 381(6656).

Gerhard Jäger. 2015. Support for linguistic macrofamilies from weighted sequence alignment. *Proceedings of the National Academy of Sciences*, 112(41):12752–12757.

Gerhard Jäger. 2018. Global-scale phylogenetic linguistic inference from lexical resources. *Scientific Data*, 5(1):1–16.

Vishnupriya Kolipakam, Fiona M. Jordan, Michael Dunn, Simon J. Greenhill, Remco Bouckaert, Russell D. Gray, and Annemarie Verkerk. 2018. A Bayesian phylogenetic study of the Dravidian language family. *Royal Society Open Science*, 5(171504):1–17.

Alexey M Kozlov, Diego Darriba, Tomáš Flouri, Benoit Morel, and Alexandros Stamatakis. 2019. Raxml-ng: a fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference. *Bioinformatics*, 35(21):4453–4455.

Mark A Larkin, Gordon Blackshields, Nigel P Brown, R Chenna, Paul A McGettigan, Hamish McWilliam, Franck Valentin, Iain M Wallace, Andreas Wilm, Rodrigo Lopez, and 1 others. 2007. Clustal w and clustal x version 2.0. *bioinformatics*, 23(21):2947–2948.

Johann-Mattis List and Robert Forkel. 2024. Lingpy. a python library for historical linguistics. With contributions by Simon Greenhill, Tiago Tresoldi, Christoph Rzymski, Gereon Kaiping, Steven Moran, Peter Bouda, Johannes Dellert, Taraka Rama, Frank Nagel, Patrick Elmer, Arne Rubehn.

Johann-Mattis List, Robert Forkel, Simon J. Greenhill, Christoph Rzymski, Johannes Englisch, and Russell D. Gray. 2023. Lexibank analysed. *Scientific Data*, 9(316):1–31. Data set.

Cédric Notredame, Desmond G Higgins, and Jaap Heringa. 2000. T-coffee: A novel method for fast and accurate multiple sequence alignment. *Journal of molecular biology*, 302(1):205–217.

S. Oskolskaya, E. Koile, and M. Robbeets. 2021. A Bayesian approach to the classification of Tungusic languages. *Diachronica*, 39(1):128–158.

Simone Pompei, Vittorio Loreto, and Francesca Tria. 2011. On the accuracy of language trees. *PLoS One*, 6(6):e20109.

Usha Nandini Raghavan, Réka Albert, and Soundar Kumara. 2007. Near linear time algorithm to detect community structures in large-scale networks. *Physical Review E—Statistical, Nonlinear, and Soft Matter Physics*, 76(3):036106.

Taraka Rama, Johann-Mattis List, Johannes Wahle, and Gerhard Jäger. 2018. Are automatic methods for cognate detection good enough for phylogenetic reconstruction in historical linguistics? In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 393–400.

Donald Ringe, Tandy Warnow, and Ann Taylor. 2002. Indo-European and computational cladistics. *Transactions of the Philological Society*, 100(1):59–129.

Laurent Sagart, Guillaume Jacques, Yunfan Lai, Robin Ryder, Valentin Thouzeau, Simon J. Greenhill, and Johann-Mattis List. 2019. Dated language phylogenies shed light on the ancestry of Sino-Tibetan. *Proceedings of the National Academy of Science of the United States of America*, 116:10317–10322.

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, and 86 others. 2023. Grambank reveals the importance of genealogical constraints on linguistic diversity and highlights the impact of language loss. *Science Advances*, 9.