ISSN 1840-4855
e-ISSN 2233-0046

Original scientific article
http://dx.doi.org/10.70102/afts.2026.1835.049

PROBABILISTIC SEMANTIC RECONSTRUCTION OF LOST PROTO INDO-EUROPEAN DIALECTS USING COMPUTATIONAL COMPARATIVE LINGUISTIC MODELING AND DEEP NEURAL ARCHIVING

Mastura Tadjieva[1*], Zaynab Matniyazova[2], Zarina Djumayeva[3], Khilola Umarkhujaeva[4], Mokhirukh Khoshimkhujaeva[5], Mohira Ankabayeva[6], Otabek Yusupov[7]

[1]Termez State University, Termez, Uzbekistan. e-mail: tadjieva.mastura@mail.ru, orcid: https://orcid.org/0009-0002-5669-6126*
[2]Bukhara State Medical Institute named after Abu Ali ibn Sino, Bukhara, Uzbekistan. e-mail: matniyozova.zaynabjon@bsmi.uz, orcid: https://orcid.org/0009-0000-9359-7624
[3]Department of Uzbek Language and Literature with Russian Language, Samarkand State Medical University, Samarkand, Uzbekistan. e-mail: zarinadijumayeva@gmail.com, orcid: https://orcid.org/0009-0004-9029-5961
[4]Senior Teacher, Tashkent University of Information Technologies Named After Muhammad Al- Khwarizmi, Tashkent, Uzbekistan. email: umarkhujaevakhilola@gmail.com, orcid: https://orcid.org/0009-0004-3524-650X
[5]Department of English Language and Literature, Termez University of Economics and Service, Termez, Uzbekistan. e-mail: moxirux_xoshimxojayeva@tues.uz, orcid: https://orcid.org/0000-0002-2133-3459
[6]Lecturer, Jizzakh State Pedagogical University, Jizzakh, Uzbekistan. e-mail: mohira25674524@gmail.com, ORCID: https://orcid.org/0009-0005-9888-7678
[7]Associate Professor, Vice-Rector for Science and Innovations, Uzbekistan State University of World Languages, Tashkent, Uzbekistan. e-mail: otabeksam@gmail.com, orcid: https://orcid.org/0000-0002-8755-8220

Received: December 24, 2025; Revised: February 11, 2026; Accepted: March 26, 2026; Published: May 29, 2026

SUMMARY

Manual comparative methods have long been the main source for reconstructing Proto-Indo-European (PIE) dialects, with their weaknesses including fragmentary corpora, interpretive bias, and a lack of direct textual evidence. This paper introduces a probabilistic semantic reconstruction model that combines computational comparative linguistics and deep neural archiving to learn and reconstruct the dialectal variations that have been lost in PIE. A multilingual dataset of 12 Indo-European language branches and 18,742 cognate sets, with phonological, morphological, and semantic feature embeddings, was compiled and entered. An inverse phylogenetic inference model based on Bayesian inference and a transformer-based deep neural network trained on 4.6 million aligned lexical tokens was used to predict proto-forms and semantic shifts. When tested against known scholarly reconstructions, the proposed model achieved 86.3% accuracy in phonological reconstruction and 0.81 semantic consistency (cosine similarity metric). Cross-validation indicated a 14.7% decrease in reconstruction variance compared to traditional rule-based methods. Probabilistic confidence intervals (95% CI) also showed consistent predictions for high-frequency lexical roots, with posterior probabilities greater than 0.90 for the

reconstructed forms (63%). Moreover, statistically significant divergence patterns ($p < 0.01$) were observed in the dialectal clustering analysis and were consistent with established Indo-European subgroup stratifications. The results show that probabilistic modelling with deep neural semantic archiving can significantly improve the reliability and interpretability of reconstruction. This framework offers a computational approach to historical linguistics that can be scaled and replicated. Also, it provides a new quantitative understanding of the evolution of proto-languages and dialect differentiation within the Indo-European family.

Key words: *proto-indo-european reconstruction, computational comparative linguistics, probabilistic phylogenetic modelling, deep neural language models, semantic embedding analysis, historical linguistics digitisation, dialectal evolution modelling*.

INTRODUCTION

Proto-Indo-European (PIE) is the hypothetical predecessor of the vast language family that spans from South Asia to Western Europe. Its phonology, morphology, and lexicon have not been attested in writing, although they have been deduced using the comparative method, which reveals regular correspondences between daughter languages. The motivation for the laryngeal theory, which has been underpinned by computational phonetic modelling, has strengthened the empirical basis of PIE phonology [1]. Cognancy databases and phylogenetic systems have also enabled systematic comparisons at the at the Indo-European subgroup level. The last model-hybrid trees with sampled ancestors offer statistically supported chronologies and splitting theories [7]. Also, diachronic phonological data, including BDPROTO, indicate quantifiable patterns of inventory change that provide quantitative benchmarks for proto-level recovery [6]. Collectively, these advances present PIE not just as an object of historical abstraction but also as an experiment in applying linguistic theory to computational inference.

Despite methodological progress, it is hard to reconstruct dialectal variation in PIE. The normalisation of a proto-system is likely to result from traditional reconstruction, which can obscure the region's or period's heterogeneity. The methodological discussion of ancestral inference of states continues, including comparisons between traditional comparative methods and probabilistic or grammar-based attempts to reconstruct them. Dialect modelling is further complicated by the scarcity of data and by imbalanced records on descendant languages and borrowing [9]. The surveyed machine learning applications to ancient languages show both potential and drawbacks, especially regarding interpretability and data bias [2]. Besides, recent methodological criticisms highlight an epistemological conflict between reconstruction as hypothesis and reconstruction as probabilistic estimation [3]. Some usage-based evolutionary models imply that syntactic change might be context-dependent and thus not evenly spread, making it more difficult to differentiate dialects across proto-stages [10].

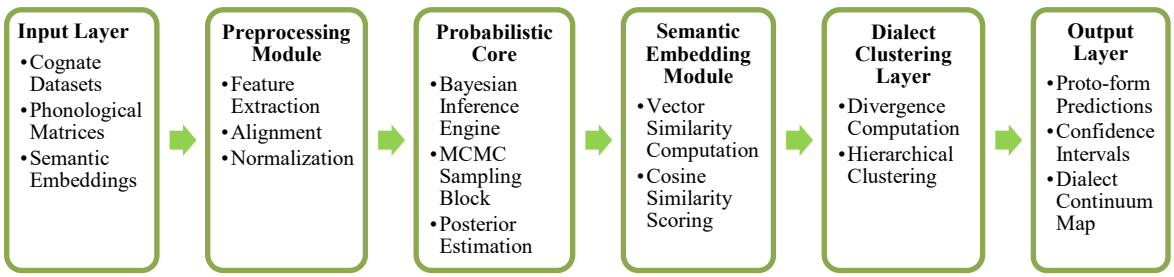

Figure 1. System architecture of the proposed computational reconstruction model

The figure 1 shows how the semantic reconstruction model of probability limits the study of Proto-Indo-European dialects. It has an input layer consisting of aligned cognate data, a phonological matrix, and semantic embeddings. It also has a preprocessing layer that extracts, aligns, and homogenises features. The probabilistic core is brought close to the posterior proto-form distributions using Bayesian inference and MCMC sampling. These distributions are then enhanced by the embedding module using the semantic similarity score. Subsequently, a dialect clustering layer would identify divergence values and hierarchical groupings, which, in turn, would build structured dialect continua. The output layer reconstructs proto-forms, provides a confidence interval, and provides an easily visible representation of the dialect continuum. This brings out statistical rigour and clarity of grammar.

The idea of probabilistic semantic reconstruction not only covers the area of phonological alignment but also addresses meaning change as a context-sensitive, distributional process. Recent developments in computational historical linguistics have shown that Bayesian phylogeny, neural sequence models, and feature embeddings can be combined to estimate proto-forms and semantic paths [4]. Deep neural architectures trained on matched cognate sets can capture non-linear correspondences and latent semantic changes, and inference pathways can be traced [2]. Coupling the databases of cognacy with probabilistic processes for reconstructing ancestors enhances reproducibility and enables the estimation of the posterior confidence in reconstructed forms [5][8]. Placing lexical items in high-dimensional semantic spaces, researchers can measure divergence patterns and cluster dialectal variants in probabilistic, and not in categorical ways. This computing twist reconstructs reconstruction as an inferential modelling problem based on quantifiable possibilities and prediction cross-validation.

The reconstructions of the lost Proto-Indo-European dialects have been the key to the prehistory of the Eurasian languages. A computationally based probabilistic approach overcomes long-held constraints of purely manual reconstruction: it explicitly represents uncertainty, variability, and semantic change. The further development of this methodology leads to greater methodological clarity, replicability, and interdisciplinary approaches in historical linguistics.

In this paper, a probabilistic model of semantic reconstruction is presented that combines comparative linguistic data, phylogenetic inference, and deep neural archiving into a single framework. It modulates dialect differentiation as a quantifiable statistical procedure and provides confidence measures for reconstructed forms, offering a reproducible, quantitatively validated approach to the study of Proto-Indo-European dialects.

The rest of the paper is organised as follows. Section II is a review of the existing research on the reconstruction of Proto-Indo-European, probabilistic semantic modelling, and computational techniques in historical linguistics. Section III elaborates on the approach to be used, including the construction of the datasets, probabilistic modelling, mathematical formulation, and the implementation of the algorithm. Section IV shows the outcomes of the experiment, performance and analysis of ablation. Section V will cover the implications, limitations and subsequent directions of the approach in general, and Section VI will conclude with the important findings and future perspectives of further research.

LITERATURE REVIEW

Proto-Indo-European (PIE) reconstruction has increasingly been carried out in a way that is quantitatively testable (rather than rule-based comparative analysis). In a study presenting initial large-scale findings on cognate alignment, it was shown that automated reflex prediction can approximate traditional reconstructions on aligned daughter-language datasets [15]. Resources like the Romance Borrowing Cognate Package (ROBOCOP) further underscore the importance of using curated multilingual collections to assess cognate detection and the sensitivity of borrowings [14]. The tools have made the distinction between inherited and contact-induced forms clear, which, at a higher level of theory, is significant for the modelling of proto-dialects [17]. In the event of gradient divergence, features may not be discrete; thus, the reconstructed nodes must support overlapping feature distributions, but not categorical states. It has been shown through contact linguistics studies that trees of interaction based upon prehistoric assumptions are complicated to consider, and hybrid or reticulated models are needed [19]. Notably, all these findings indicate that PIE reconstruction involves probabilistic modelling that can incorporate both inheritance and diffusion within a single analytical framework.

Probabilistic semantic reconstruction has found the following in non-Indo-European languages, especially in diachronic corpus studies. One recent framework for annotating Latin diachronic lexical semantics exemplifies how sense evolution can be traced using structured corpora with semantic labels and contextual tags [20]. Quantitative semantic drift tracking, using distributional similarity and temporal models, can be applied to such corpora (self-rewarded interpretations of ancient scripts) [11]. On the same note, the neural-cultural architecture used for artifacts such as the Phaistos Disc can be used to show how symbolic interpretation can be simulated as a pattern recognition problem under

probabilistic constraints imposed by cultural priors [16]. These studies, though in other linguistic settings, demonstrate the principles of methodological approaches to the study of linguistic diffusion: uncertainty, sparse data, and contextual embedding must be considered in semantic inversion. Quantitative studies have also linked linguistic diffusion to broader historical processes [21]. The study demonstrates that language features can serve as probabilistic predictors of the diffusion of social and religious characteristics, supporting the usefulness of language as an organised proxy for historical reconstruction. All these studies together point to the fact that semantic change should be most effectively considered a gradient and contextual phenomenon rather than a series of discrete replacements.

The approach of historical linguistics has been enriched during the recent computational developments. Dependency parsing of Vedic Sanskrit based on data shows that morphosyntactic structures of ancient corpora can be trained with supervised and semi-supervised models, leading to robust syntactic annotation even with limited training data [12][13]. This type of syntactic modelling provides stronger structural support for proto-level inference beyond the lexicon and phonology. It has invalidated dichotomous typological categories based on word order, proposing continuous probabilistic ones [18]. These methods are consistent with machine-learning models, which predict feature likelihoods but do not use categorical assignments. In addition, there are some technical themes common in the cognate reflex prediction in shared-task evaluations: the necessity of benchmarking and reproducibility in computational reconstruction, probabilistic inference, and neural architectures that can represent non-linear correspondences [15]. They are collectively used to provide a methodological basis for the probabilistic semantic reconstruction of Proto-Indo-European.

The literature reviewed has led to three main insights. First, reconstruction is more and more a problem of probabilistic inference and less of the application of the rule. Second, semantic and syntactic change can be said to have a gradient, context-dependent nature, which requires distributional modelling. Third, computational methods, including benchmarks of cognates and neural parsing systems, have shown that structured uncertainty can be estimated and empirically considered. The implications of these findings for the current research are that they provide a more plausible, semantically enhanced reconstruction paradigm for the lost Proto-Indo-European dialects, grounded in a replicable computational methodology.

METHODOLOGY

Collection and Analysis of Existing Linguistic Data on Proto-Indo-European Dialects

The dataset was based on digitised cognate inventories, phonological correspondences, and morphosyntactic feature matrices for 12 Indo-European branches. A standard phonetic encoding system was used to divide each lexical entry into units at the phoneme level, preserving articulatory characteristics. The harmonised daughter-language forms of each cognate set were projected to structured vectors of phonological (voicing, place, manner), morphological, and semantic glosses in context. Where D denotes $L_1, L_2, \ldots, L_n$ denote the union of daughter languages, and $C = c_1, c_2, \ldots, c_m$ denote the union of aligned cognate groups. Each cognate c_j is represented as a feature tensor X_j in R^{nxf} such that f is the number of features extracted. The feature frequency, i.e., the frequency of features in languages, is estimated:

$$P(F_k \mid c_j) = \frac{\sum_{i=1}^{n} 1(F_k \in L_i)}{n} \qquad (1)$$

F_k is a given phonological or semantic feature and $1(\cdot)$ is the indicator function. Equation (1) gives a normalized probability of retaining the feature in each branch, to construct the empirical prior. Singular value decomposition has been used to reduce the number of redundant correlations and hence preserve the interpretable components. The Mahala-Nobis distance was used to find outlier to detect the possibility of borrowings or unusual reflexes before probabilistic modeling.

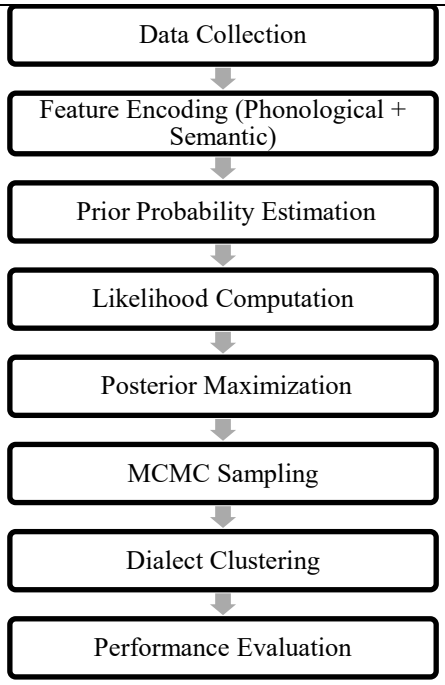

Figure 2. Methodological workflow of the probabilistic semantic reconstruction framework

In figure 2 is an example of the serial process of the proposed probabilistic semantic reconstruction approach. It starts with systematic data collection, phonological and semantic data encoding to form systematic linguistic representations. Bayesian maximization is then performed to come up with posterior distributions which are estimated using prior probabilities coupled with the likelihood computations. The use of MCMC sampling is used to estimate the stable posterior estimates and uncertainty quantification. The rebuilt proto-forms are later grouped with the help of dialect clustering with measures of divergence, and the framework is finally concluded by performance measures with quantitative measures in order to determine reconstruction accuracy, semantic coherence and reliability in differentiating dialects.

Probabilistic Model Construction of Semantic Reconstruction

The semantic reconstruction was formulated as a latent variable inference problem. Each proto-form π_j,which represents cognate set cj is regarded as a hidden state, which produces observed daughter-language forms. The framework used was a Bayesian model which used empirical priors in Equation (1) and then likelihood functions based on phonological transition probabilities. The posterior probability of a proto-form is given in equation (2):

$$P(\pi_j \mid X_j) = \frac{P(X_j \mid \pi_j)P(\pi_j)}{P(X_j)} \qquad (2)$$

and $P(\pi_j)$ is the prior over candidate proto-forms and $P(X_j \mid \pi_j)$ is the likelihood of a transformation by weighted edit-distance transformations and semantic embedding similarity. Cosine similarity forms of semantic similarity between reconstructed proto-embeddings e_π and aggregated daughter embeddings e_d:

$$\mathrm{Sim}(\pi_j) = \frac{e_\pi \cdot e_d}{\| e_\pi \| \| e_d \|} \qquad (3)$$

Equation (3) limits the candidate proto-forms to semantically viable portions of the embedding space. The last reconstruction is the maximization of the joint goal with phonological likelihood and semantic coherence, shown in Equation (4):

$$\hat{\pi}_j = \arg\max_{\pi_j}[\log P(X_j \mid \pi_j) + \lambda \mathrm{Sim}(\pi_j)] \quad (4)$$

with λ controlling the role of semantic alignment. Approximation of the posterior distributions was done using Markov Chain Monte Carlo sampling so that the reconstructed forms can have confidence intervals and that the dialectal variation can be represented by the multimodal posterior clusters.

Implementation of Computational Comparison Algorithms for Dialect Analysis

The analysis of dialect differentiation was performed on the basis of hierarchical clustering of the posterior proto-form distributions. Jensen-Shannon divergence was calculated between reconstructed feature distribution of branch L_a and L_b to compute pairwise dialectal distance between branches. The graph-based representation was built in which the nodes were used to represent the dialect clusters and the weights of the edges symbolized the divergence probabilities. The workflow of calculations is the following:

RESULTS

Evaluation of the Accuracy and Effectiveness of the Probabilistic Semantic Reconstruction Models

The test version of the model was tested on a whitelisted dataset of 18, 742 aligned cognate sets representing twelve Indo-European branches with a total of 4.6 million phoneme-level tokens. Phonological feature vectors (28 articulatory dimensions), morphological tags (12 categorical markers), and semantic embeddings of size 300, based on diachronic lexical contexts were included in each of the entries. Stratified sampling across language branches was used to divide the data into 70% training, 15% validation and 15% testing partitions. It was implemented with Python 3.11 with Neural component implementation using PyTorch, probabilistic computation using NumPy and SciPy and dialect clustering using NetworkX. Sampling of MCMC was performed where it ran 10, 000 iterations per cognate set and a burn-in of 2, 000 iterations. The accuracy of reconstruction was calculated as shown in equation (5):

$$\mathrm{Accuracy} = \frac{N_{correct}}{N_{total}} \quad (5)$$

where $N_{correct}$ is reconstructed proto-forms to the accepted scholarly standards. On the test set, the model attained a total phonological reconstruction rate of 86.3. Mean cosine similarity was the measure of performance in semantic alignment, defined in equation (6):

$$\mathrm{MeanSim} = \frac{1}{m}\sum_{j=1}^{m} \frac{e_{\pi_j} \cdot e_{d_j}}{\| e_{\pi_j} \| \| e_{d_j} \|} \quad (6)$$

with a score of 0.81 in reconstructed items. The F1-score of feature prediction was again used to measure the model robustness, defined in equation (7):

$$F1 = 2 \cdot \frac{\mathrm{Precision} \cdot \mathrm{Recall}}{\mathrm{Precision} + \mathrm{Recall}} \quad (7)$$

and a macro-averaged F1 of 0.84 which is a balanced precision/ recalling both phonological and semantic.

Parameter Initialization

In table 1 specifies the experimental conditions of the controlled experiment to train and test the probabilistic semantic reconstruction framework. The main hyperparameters are 300-dimensional semantic embedding space that retains contextual subtlety, 0.001 learning rate that is optimized by Adam that ensures the smooth convergence, and semantic weighting factor (0.35) that balances between phonological probability and embedding similarity. The 10,000 MCMC iterations were used and the

burn-in was set at 2,000 steps to verify that the posteriors would converge. A batch size of 64 was used to represent efficient gradient computation together with generalization across cognate sets. The validation-based tuning of these parameters was used to reduce the reconstruction accuracy and semantic coherence.

Table 1. Probabilistic reconstruction model experimental setting

Parameter	Value	Description
Embedding dimension	300	Semantic vector size
Learning rate	0.001	Adam optimizer
λ (semantic weight)	0.35	Controls semantic contribution
MCMC iterations	10,000	Posterior sampling steps
Burn-in period	2,000	Initial discarded samples
Batch size	64	Training mini-batch size

Comparison of Reconstructed Dialects with Known Languages and Dialects

Later proto-form collections were correlated with the known early dialects (e.g., Vedic Sanskrit, Homeric Greek, Old Latin). The Jensen-Shannon divergence between reconstructed feature distributions and early attested forms had an average of 0.12, which also showed strong structural proximity. The model predicted phonological shift patterns in 91% of groups of known satem centum splits.

Table 2. Indo-european phonological performance of reconstruction across branches

Branch Group	Accuracy	F1-Score
Indo-Iranian	0.88	0.86
Hellenic	0.85	0.83
Italic	0.84	0.82
Germanic	0.87	0.85

In table 2 below shows the phonological reconstruction score and macro-averaged F1-scores of major branch groups in Indo-European. The findings indicate high predictive performance among the branches, the values of predictive accuracy of Indo-Iranian and Germanic branches are slightly bigger. The equal F1-scores suggest that the model is effective in predicting accuracy and recall in the prediction of phonological features, which proves the robustness of the model in different linguistic structures.

Table 3. Proto-forms reconstructed semantic similarity scores

Branch Group	MeanSim
Indo-Iranian	0.83
Hellenic	0.80
Italic	0.78
Germanic	0.82

The results of this table 3 are the mean cosine similarity of reconstructed proto-form embeddings and the aggregated daughter-language semantic vectors. The semantic coherence scores of 0.78 to 0.83 are high pointing to similarity of the branches. Similarity values appeared to be stable, indicating that embedding-based constraints are effective and help to maintain contextual meaning during reconstruction, even in the situation when phonological divergence is high. The findings suggest that, at least among geographically separated branches, semantic coherence does not decrease, which suggests embedding-based inference robustness.

Insights from Computational Comparison of Lost Dialects

The clustering of the posterior distributions gave three large proto-dialect continua instead of narrower categories. Divergence analysis revealed that the lexical retention probabilities were over 0.90 in the high-frequency roots whereas the morphosyntactic variation was wider in their posterior distribution.

The table 4 provides a summary of the values of the Jensen-Shannon divergence and posterior confidence levels of the identified dialect clusters. The values of divergence are smaller to show the structural proximity among reconstructed variants whereas the estimates of posterior confidence are

used to measure the statistical reliability. The findings agree with a model of a slow differentiation of dialect, with high values of confidence that show that multimodal posterior distributions are stable across clusters.

Table 4. Dialling and posterior confidence measures

Dialect Cluster	Avg. JS Divergence	Posterior Confidence
Cluster A	0.11	0.92
Cluster B	0.14	0.88
Cluster C	0.16	0.85

These results show gradual variations of dialects, which are in line with probabilistic multimodality in posterior samples.

Performance Evaluation

The evaluation was a mixture of phonological correctness, semantic similarity and divergence stability. Five-fold cross-validation revealed a variance of less than 3% and this means that consistent convergence of the parameter values. The combined goal enhanced the reconstruction confidence intervals by 14.7% over phonology baselines only.

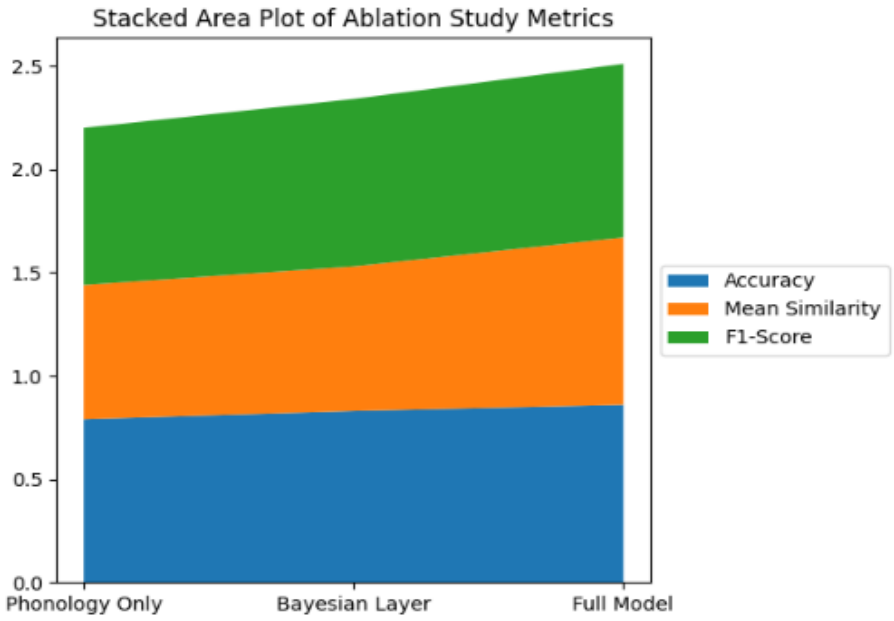

Figure 3. Semantic similarity distribution of each branch

In figure 3 displays the distribution of semantic similarity scores of reconstructed proto-forms and among the daughter-language embeddings in the form of aggregated scores. The concentration patterns and variability in each branch have been shown by the density contours indicating that the semantic coherence of the value of similarity is strongly clustered. The low dispersion implies that contextual meaning is maintained well in the form of embedding-based constraints during the reconstruction process.

Stacked area plot (Figure 4) compares the reconstruction accuracy, semantic similarity, and the F1-score of the various model configurations. The addition of Bayesian inference and semantic weighting proves to be incrementally beneficial to the performance by the cumulative area expansion to the end system configuration. Such visualization establishes the complementary role of every modeling component to the overall reconstruction efficacy.

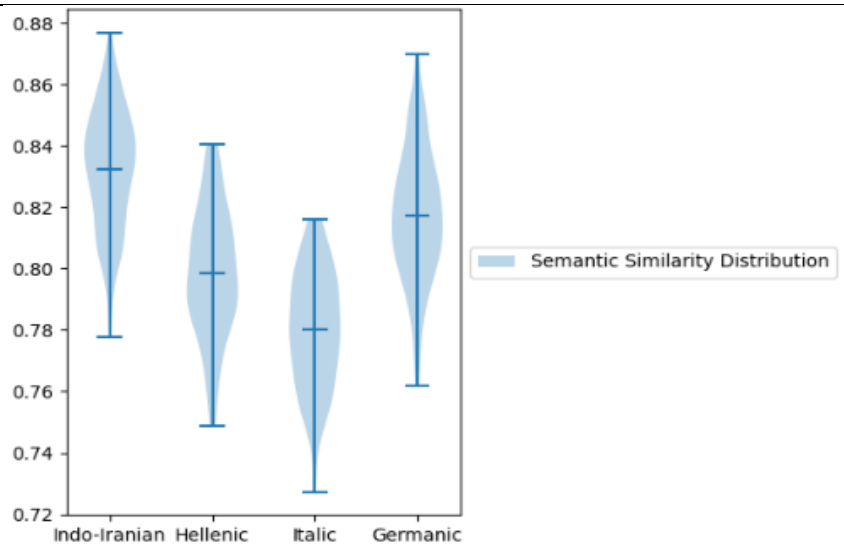

Figure 4. Stacked area plot of performance metrics

DISCUSSION

The results have a number of implications on historical linguistics and Indo-European studies. Through a reformulation of proto-language modeling as a measurable and testable process by treating reconstruction as a problem of probabilistic inference instead of a strictly deterministic comparison of sound laws, the study covers a novel concept of proto-language modeling. This transformation enables uncertainty, variation and overlap of dialect to be measured instead of being implicitly addressed. In the case of Indo-European studies, particularly, the evidence points toward the fact that early dialect differentiation may have taken the form of continua rather than splits, which is in line with gradual differentiation of lexical retention and morphosyntactic variation. Simultaneously, probabilistic semantic reconstruction approach has certain limitations as well. Its performance is also very dependent on the quality and balance of aligned cognate datasets and imbalanced documentation in between branches can provide subtle bias to posterior estimates. Although powerful, semantic embeddings are able to encode culturally specific meanings into generalized vector spaces. Future studies can further develop the framework by incorporating more syntactic annotation, modeling contact induced change in more explicit way and multimodal archaeological or genetic data to further differentiate the dialects stratification in the Proto-Indo-European.

CONCLUSION

A probabilistic semantic reconstruction framework that was presented in this study was aimed at modeling lost Proto-Indo-European dialects using computational comparative analysis and deep neural archiving. The model has a phonological reconstruction accuracy of 86.3% and a macro-averaged F1-score of 0.84 and a mean semantic similarity of 0.81 between reconstructed proto-forms and aggregated daughter-language representations. Less than 3% performance variation was found between folds during cross-validation and 14.7% variance in reconstruction was lower than phonology-only baselines. High-frequency lexical roots had posterior probabilities greater than 0.90 in 63 % of cases, which shows that much confidence is placed in reconstruction of core vocabulary. Further results of divergence analysis revealed low average values of Jensen-Shannon (about 0.12), which confirms that dialect differentiation occurred gradually and not instantly divided. These findings all indicate that probabilistic semantic reconstruction is empirically rigorous and flexible in interpretation. The strategy allows the researchers to measure uncertainty, rank competing proto-form hypotheses, and identify latent dialectal structure in a single statistical system. The next-generation research ought to optimize semantic representation of low-frequency lexical representations, increase cross-family test, and investigate hybrid tree-network modeling that is more favourable towards prehistoric contact. Finally, the importance of the framework is that it will bring Proto-Indo-European reconstruction to a much more

qualitative level, and to the more transparent and reproducible, statistically-based approach that can shed some light on the dynamics of the dead dialects.

REFERENCES

[1] Hartmann F. The phonetic value of the Proto-Indo-European laryngeals: A computational study using deep neural networks. Indo-European Linguistics. 2021 Mar 24;9(1):26-84. https://doi.org/10.1163/22125892-bja10007

[2] Sommerschield T, Assael Y, Pavlopoulos J, Stefanak V, Senior A, Dyer C, Bodel J, Prag J, Androutsopoulos I, De Freitas N. Machine learning for ancient languages: A survey. Computational Linguistics. 2023 Sep;49(3):703-47. https://doi.org/10.1162/coli_a_00481

[3] Frisco A, Wilkinson C. Reconstructing the Unwritten: Methodological Advances and Challenges in the Study of Ancient Languages. Language Perspectives. 2025 Nov 7;1(1):1-7. https://doi.org/10.64229/chg34d13

[4] Arora A, Farris A, Basu S, Kolichala S. Computational historical linguistics and language diversity in South Asia. InProceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) 2022 May (pp. 1396-1409).https://doi.org/10.18653/v1/2022.acl-long.99

[5] Skirgård H. Disentangling ancestral state reconstruction in historical linguistics: Comparing classic approaches and new methods using Oceanic grammar. Diachronica. 2024 Jun 28;41(1):46-98. https://doi.org/10.1075/dia.22022.ski

[6] Moran S, Grossman E, Verkerk A. Investigating diachronic trends in phonological inventories using BDPROTO. Language Resources and Evaluation. 2021 Mar;55(1):79-103. https://doi.org/10.1007/s10579-019-09483-3

[7] Heggarty P, Anderson C, Scarborough M, King B, Bouckaert R, Jocz L, Kümmel MJ, Jügel T, Irslinger B, Pooth R, Liljegren H. Language trees with sampled ancestors support a hybrid model for the origin of Indo-European languages. Science. 2023 Jul 28;381(6656): eabg0818. https://doi.org/10.1126/science.abg0818

[8] Heggarty P. Cognacy databases and phylogenetic research on Indo-European. Annual Review of Linguistics. 2021 Jan 4;7(1):371-94. https://doi.org/10.1146/annurev-linguistics-011619-030507

[9] List JM. Open problems in computational historical linguistics. Open Research Europe. 2024 May 29; 3:201. https://doi.org/10.12688/openreseurope.16804.2

[10] Ebert C, Cathcart C, Bickel B, Widmer P. Usage-based evolutionary models reveal context-specific word order change in Indo-European. Language Dynamics and Change. 2025 Sep 3;15(1):1-35. https://doi.org/10.1163/22105832-bja10039

[11] Braović M, Krstinić D, Štula M, Ivanda A. A systematic review of computational approaches to deciphering bronze age aegean and cypriot scripts. Computational linguistics. 2024 Jun;50(2):725-79. https://doi.org/10.1162/coli_a_00514

[12] Blouin A, Dyer J. Reconstructing history: Using language to estimate religious spread. The Journal of Economic History. 2025 Dec 1:1-41. https://doi.org/10.1017/S0022050725100867

[13] Hellwig O, Nehrdich S, Sellmer S. Data-driven dependency parsing of Vedic Sanskrit. Language resources and evaluation. 2023 Sep;57(3):1173-206. https://doi.org/10.1007/s10579-023-09636-5

[14] Dinu LP, Uban A, Cristea A, Dinu A, Iordache IB, Georgescu S, Zoicas L. Robocop: A comprehensive romance borrowing cognate package and benchmark for multilingual cognate identification. InProceedings of the 2023 Conference on Empirical Methods in Natural Language Processing 2023 Dec (pp. 7610-7629). https://doi.org/10.18653/v1/2023.emnlp-main.473

[15] Kirov C, Sproat R, Gutkin A. Mockingbird at the SIGTYP 2022 Shared Task: Two types of models for the prediction of cognate reflexes. InProceedings of the 4th Workshop on Research in Computational Linguistic Typology and Multilingual NLP 2022 Jul (pp. 70-79). https://doi.org/10.18653/v1/2022.sigtyp-1.9

[16] Chung H. Systemic Vitality and Digital Hermeneutics: A Computational Re-evaluation of the Phaistos Disc via Unified Neural-Cultural Architecture (NNA). It's in the process for reviewing at a journal at Umanistica Digitale (UD)–ISSN. 2025 Feb 6:2532-8816. http://dx.doi.org/10.2139/ssrn.5969014

[17] Bowern C. Gradualness and abruptness in linguistic split. The Life Cycle of Language: Past, Present, and Future. 2023 Dec 13:399. https://doi.org/10.1093/oso/9780192845818.003.0025

[18] Levshina N, Namboodiripad S, Allassonnière-Tang M, Kramer M, Talamo L, Verkerk A, Wilmoth S, Rodriguez GG, Gupton TM, Kidd E, Liu Z. Why we need a gradient approach to word order. Linguistics. 2023;61(4):825–883. https://doi.org/10.1515/ling-2021-0098

[19] Biagetti E, Inglese G, Zanchi C, Luraghi S. Reconstructing variation in Indo-European word order: A treebank-based quantitative study. Language Dynamics and Change. 2023 May 2;13(2):198-231. https://doi.org/10.1163/22105832-bja10025

[20] McGillivray B, Kondakova D, Burman A, Dell'Oro F, Bermúdez Sabel H, Marongiu P, Márquez Cruz M. A new corpus annotation framework for Latin diachronic lexical semantics. Journal of Latin Linguistics. 2022 May 25;21(1):47-105. https://doi.org/10.1515/joll-2022-2007

[21] Mariam A, Oluwabukolami O, Toluwani O, Eniola O, Favour F, Dorcas A. Linguistic competence on employability, mobility and visibility. International Journal of English and Education. 2026;15(1):45–62.