Analyzing Pre-Indo-European Theory of Etruscan Language Origins Using Topological Data Analysis

New Mexico

Supercomputing Challenge

Final Report

April 2, 2025

Welch Homeschool

Team Members:

Helena Welch

Teacher:

Cindy Welch

Project Mentor:

Paul Welch

Executive Summary

High-dimensional data is often difficult to analyze because of the exponential growth of the size of the space in which the data lives as the dimension increases. [16, 3] One example of high-dimensional data comes from language, which contains many different characteristics (dimensions) with which it can be quantified, but not always sufficient data to detect patterns in it. This is especially true for ancient languages, as there is a sparsity of texts from which to draw. [6]

The ancient Etruscan language is currently classified as a non-Indo-European isolate. However, the Etruscans lived in an Indo-European-speaking region and appear to be genetically related to Indo-Europeans. [12, 28] This study aims to bring a quantitative measure, topological data analysis (TDA), to ongoing investigations of Etruscan to more concretely determine Etruscan's similarity to different Indo-European languages. Phonetic patterns in a specific word list translated into different languages by large-language models are encoded, and the distance between two given phonemes based on this encoding is calculated. Results indicate that Sanskrit has the highest correlation to Etruscan. Etruscan appears similar to older Indo-European languages and thus may be older than neighboring languages, explaining its uniqueness compared to Indo-European languages that developed later in time.

Introduction

The Questions Behind Etruscan as a Non-Indo-European Language

Etruscan was spoken in the Etrurian region of Italy around 800 to 100 BC (see Figure 1). [12] With more than 13,000 examples of Etruscan text, we have enough data to understand a fair amount of vocabulary but not to concretely determine the origins of the language. [12] In addition, while much about the language is unknown, linguists have been able to reconstruct in part the sounds of the language because the Etruscan alphabet was borrowed from that of the Euboean Greeks and subsequently passed on to the Romans. [4] Thus, the phonology of the language is fortunately available for use in this study.

The current reigning theory posits that Etruscan was one of the few Pre-Indo-European languages not displaced by the arrival of the Indo-European language family. The only languages that seem most certainly to be related to Etruscan are the obscure Rhaetic and Lemnian languages, of which only a handful of texts are extant. [37] Despite this, however, a recent genetic analysis indicates that the Etruscans were indigenous to Italy and had very similar genes to those of the Romans. [28] If this is the case, then why is their language believed to be so different from that of their Indo-European-speaking relatives?

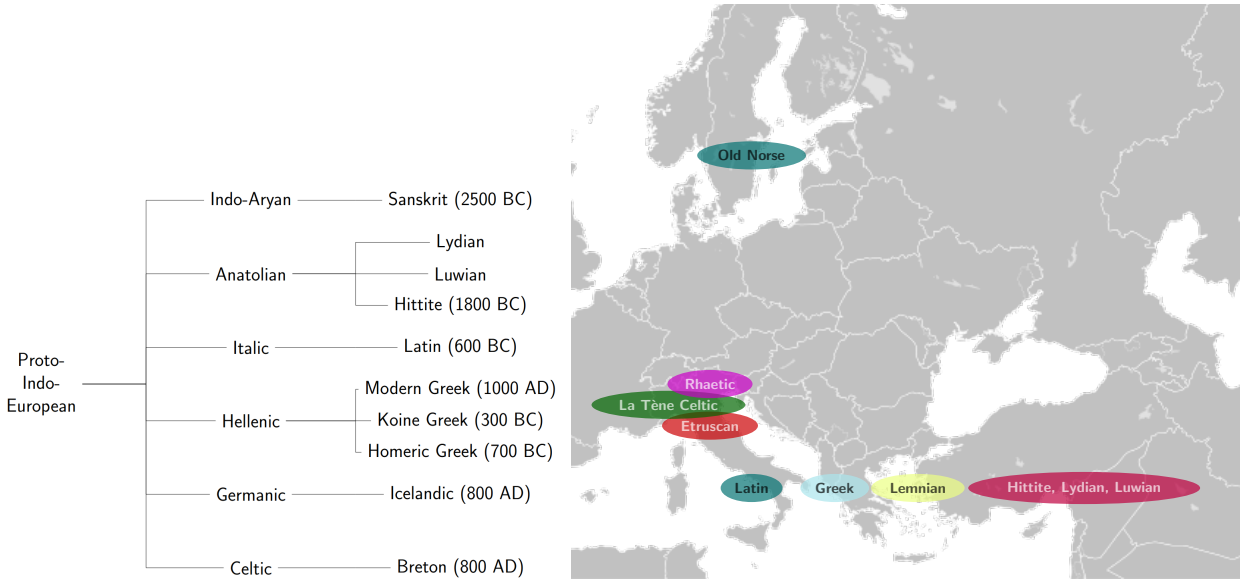

Figure 1: Etruria compared to regions where other ancient languages were spoken. [18]

Other Theories

While the Pre-Indo-European theory is currently most popular, other studies have suggested alternatives. The Anatolian theory originates from writings of the ancient historian Herodotus, which claim that the Etruscans were Lydians who migrated from Anatolia (modern-day Turkey) to Italy. [30] If this is true, Etruscan should exhibit similar linguistic properties to Hittite, a better-recorded relative of Lydian. Others suggest Etruscan was a Greek creole language made up of a mixture of the Greek dialects. The enigma of Etruscan's origin runs so deep that even hypothetical links between Old Norse, the Celtic languages, and Aryan languages such as Sanskrit have been suggested. [32, 13, 7, 15]

Purpose of Project

This project compares the phonology of ancient and contemporary Indo-European languages to that of Etruscan using topological data analysis (TDA), with the aim of discovering which Indo-European languages are most closely related to Etruscan. While statistics aims to fit data points to lines or other geometries, TDA quantifies the distance between data points in a high-dimensional space using topological structures such as n-dimensional holes. [31, 35, 33] This means that it captures higher-level information about the inherent structure of data rather than looking for specific patterns in it that might not exist. [31, 38] Because of this, previous studies have used TDA to detect similarities between languages based on their phonological and grammatical structures. [26, 38] It is known to perform well in comparison to other techniques because of its insensitivity to sparsity of data and noise, which could make it particularly useful in analyzing poorly-documented ancient languages. [38, 35, 33]

Prior Works

While TDA appears to be a novel approach to implement comparative linguistics, two studies that develop a methodology for doing so for different aspects of linguistics are cited here. The first uses TDA to compare grammatic parameters between modern languages [26], while the second proposes an algorithm for comparing the phonetics of languages [38]. The methodology of this study is largely based on the latter's work.

Wolfram's study (2017) performed TDA for large families of modern languages using a list of 200 words and, although laying an extensive groundwork for future work, ultimately determined that a larger data set was needed to prove more conclusive results. In contrast, this study chooses a set of 1700 words and performs Wolfram's (2017) methodology on an ancient language whose ancestry is uncertain. In order to obtain a larger dataset than Wolfram (2017) could acquire, this study uses large-language models (LLMs) to translate a single word list into different languages.

Methods

Data Collection and Preparation

A word list of 1700 Etruscan words and their English translations is drawn from McCallister [1999], and each phoneme in an Etruscan word is then converted to its corresponding International Phonetic Alphabet (IPA) character using conversions provided by Rix [2008] and Ager [Omniglot. 2023]. IPA is a system for encoding phonemes spoken in different languages such that the list of sounds is universal across all languages.

Having acquired an Etruscan word list with its corresponding English translation and converted it to IPA, translations of the word list into other languages are then obtained through the use of large-language models (LLMs). LLMs are chosen to overcome the roadblock of limited available translation data that previous studies met. [38] Each of five LLMs is run from the Ollama server [24] on one of two desktop computers (see Table 1), depending on the size of the models. In addition, a relatively larger model, ChatGPT4o, is run from the OpenAI server. [25] Its running time and cleanliness of the output are notably better.

Each LLM is passed a message communicating its role as an assistant, as well as the prompt, "I will give you a number of words to translate into [a language]. Provide the International Phonetic Alphabet. Do not provide any notes or commentary. Use the format: English: [language]: /IPA/." The list of 1700 words in English is then read from a file and passed in batches of five words as input to the model. Output is then written to a file and parsed to obtain a list of the translation in IPA. While all LLMs are fairly consistent in providing output in the correct format, manual parsing of certain sections in models mistral-small and nous-hermes is required. In addition, all

model	parameters	vocabulary size
ChatGPT4o [34]	200 billion	175k
Mistral-large [20]	123 billion	128k
Mistral-small [22]	22 billion	32k
Command-r [10]	35 billion	128k
Nous-hermes [36]	34 billion	64k
Mistral-nemo [21]	12 billion	128k

Table 1: Characteristics of the large-language models used to translate the Etruscan word list into other languages.

IPA characters outputted across all LLMs are researched, and outdated or nonstandard symbols are replaced to ensure that all translations of the word list conform to the same IPA encoding. Ultimately, this results in an IPA character list of 34 symbols. Etruscan phonology consists of 21 of those characters. [29]

Following the procedure designed by Wolfram 2017, the contextual relations between different IPA characters within a given language are quantified by creating a list of the IPA characters that come before and after each IPA character in a word. This is known as the context list. For example, the context list of *ū* for the English IPA word list *Sūt, St, sūn, rn* would consist of (*S,t*),(*s,n*), and the context list of would be (*S,t*),(*r,n*). The cosine similarity C between two such characters for a given language is then calculated, where S is number of contexts two characters share, N_1 and N_2 are the length of each character's context list, and C is defined as:

$$C = \frac{S}{\sqrt{N_1} * \sqrt{N_2}} \tag{1}$$

This quantifies how irreplaceable one phoneme is with another within a given language. [38] One would expect, for example, that two vowels would have a higher cosine similarity than a vowel and a consonant.

This information can be encoded in a matrix, where each axis is the set of IPA characters and

each value is the cosine similarity between any two IPA characters in a given language. [38] As such, the matrix should be a symmetric, square matrix with values of 1 along the diagonal, as the cosine similarity between two identical IPA characters is 1. A visualization for Etruscan's cosine similarity matrix can be seen in Figure 2, where characters are arranged by vowels (top left corner) and consonants (bottom right). The lighter the color, the more two characters have in common. Dark rows and columns indicate that a given character does not appear in a language.

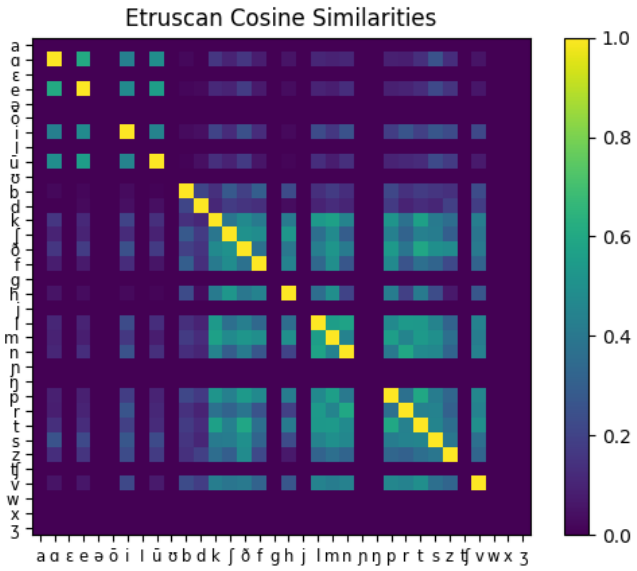

Figure 2: Colorplot for Etruscan's cosine similarity matrix created in matplotlib. [14]

Having quantified the relation between every two IPA characters in a given language, each character is embedded in a coordinate system such that its coordinate axis is the set of cosine similarities between it and each other IPA character. Thus, this point cloud lives in a space with the same number of dimensions as there are IPA characters, which, in this case, is a 34-dimensional space. If an IPA character does not occur in the word list translated into a given language, its cosine similarity with each other character is set to 0. The particular TDA used in this study, persistent homology, is then performed on the point cloud.

Persistent Homology

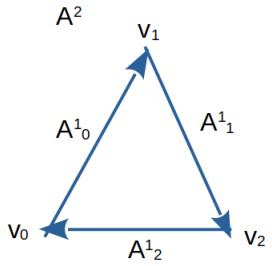

Figure 3: Example 2-simplex.

Persistent homology is performed using the Scikit-TDA python library Ripser. [31] Persistent homology is a type of topological data analysis that draws nth-dimensional balls, n-balls, around a point cloud in an nth-dimensional space. [38] At a set radius of the n-ball, all n-balls that overlap are connected by an n-simplex A^n, where an A^n is the simplest geometry determined by $(n+1)$ connected points in the Euclidean space R^n. [38] For example, 0-simplex A^0 is 1 point, 1-simplex A^1 is 2 connected points (a line segment), and 2-simplex A^2 is three connected points (a triangle).

An n-chain is a sum of n-simplices. It can be described by its boundary δ:

$$\delta(\mathrm{A}^n) = \mathrm{A}_0^{(n-1)} + \mathrm{A}_1^{(n-1)} + \ldots + \mathrm{A}_n^{(n-1)} \tag{2}$$

for n-simplices A^n.

For example, the boundary of an n-chain that contains one simplex, a 2-simplex shown in Figure 3, is

$$\delta(\mathrm{A}^2) = \mathrm{A}_0^{(1)} + \mathrm{A}_1^{(1)} + \mathrm{A}_2^{(1)} \tag{3}$$

For this 2-simplex, the boundary of the boundary is defined as:

$$\delta(\delta(\mathrm{A}^2)) = \delta(\mathrm{A}_0^1 + \mathrm{A}_1^1 + \mathrm{A}_2^1)$$

$$= \delta(\mathrm{A}_0^1) + \delta(\mathrm{A}_1^1) + \delta(\mathrm{A}_2^1)$$
$$= [(\mathrm{v}_1) - (\mathrm{v}_0)] + [(\mathrm{v}_2) - (\mathrm{v}_1)] + [(\mathrm{v}_0) - (\mathrm{v}_2)] = 0$$

(4)

Because of the 2-simplex's orientation, the boundary of each edge (1-simplex) is defined as the difference of its endpoints (the vertices of the 2-simplex). Summing the edges causes vertices to cancel, leaving $\delta(\delta(\mathrm{A}^2)) = 0$. This tells us that the vertices of the 2-simplex are connected. This generalizes to higher-dimensional simplices.

The homology group of the nth dimension, then, is the set of all n-chains that do not cause vertices to cancel. Thus, it describes groups of simplices that are unconnected and have $\delta(\delta(\mathrm{A}^n)) \neq 0$:

$$H_n = \frac{Z_n}{B_n} \tag{5}$$

where Z_n is the set of all n-chains and $B_n \in Z_n$ that have $\delta(\delta(\mathrm{A}^n)) = 0$. [23, 8]

Essentially, H_n measures which points are not connected at a certain radius of the n-balls through calculation of nth dimension holes, where n is the dimension of the simplex. [31] Since a 0-dimensional hole describes the disconnected parts of a 0-simplex, the H_0 homology group is the set of components that are not connected to each other. Each given component, then, contains all points in the 34-dimensional space that are connected by n-balls at a given radius. Each component is known as a cluster.

H_1 measures 2-dimensional holes, and H_2 measures cavitites, known as voids. Note that in the case of language studies, topological structures of a higher order than clusters and holes are often noise and, thus, will not be considered in this study. [38, 27]

Ripser encodes relations between simplices of different dimensions in a boundary matrix,

where each row is an n-simplex and each column represents an $(n+1)$ simplex. If the n-simplex is part of the $(n+1)$-simplex's boundary, it is encoded as a non-zero value in the matrix. [31]

Persistence Diagrams

The Persim library within Scikit-TDA [31] is then used to generate persistence diagrams describing results of the TDA. Figure 4 gives two such graphs, one for German and one for Dutch. These diagrams display the radius of the n-balls at which each cluster or hole is "born" (when it first appears) or "dies" (when it is consumed by another topological structure). Each axis describes the birth or death radius of a topological structure. All clusters (H_0) should fall on a vertical line, since all are born at radius $r = 0$, and it follows that only one will remain as r approaches ∞.

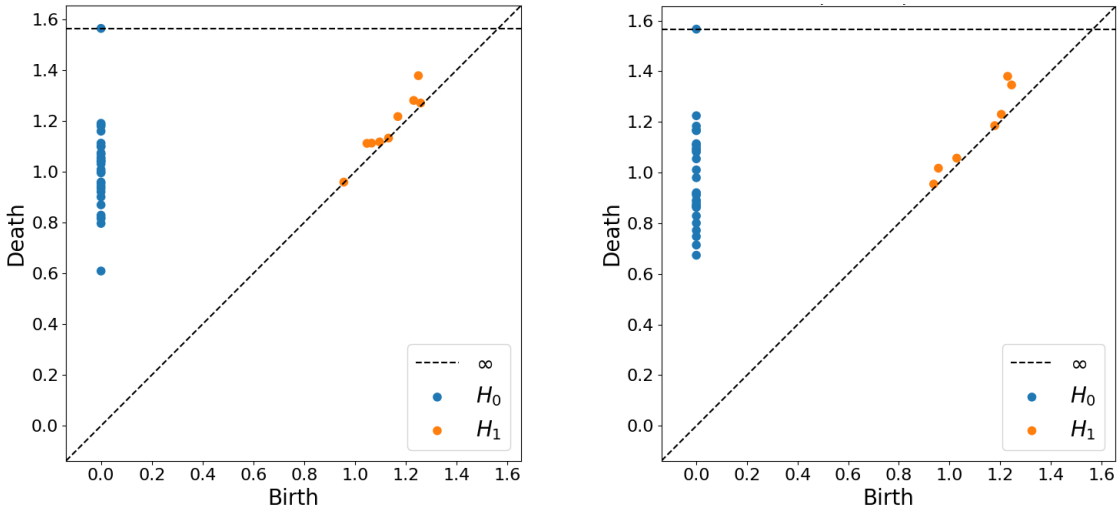

Figure 4: Persistence diagrams for German (left) and Dutch (right); input data for TDA were drawn from mistral-small translations.

Holes (H_1, shown in orange) tend to be near the diagonal, or when birth radius is equal to death radius. The farther a hole is from the diagonal, the larger the radius must become before clusters are joined and the hole disappears. Linguistically, this indicates how similar various IPA characters are to each other, based on how often they appear in the same context in a given language's word list.

Comparison Methods for Persistence Diagrams

After generating persistence diagrams, Persim compares those for Etruscan and eight other lan-

guages. Thus far, these languages consist of Latin, Breton, Koine Greek, Homeric Greek, Modern Greek, Icelandic, Hittite, and Sanskrit. Languages were chosen based on the time and proximity to Etruscan (see Figure 1), as well as availability of LLMs to accurately provide words from their vocabulary. (Icelandic, for example, is known to be linguistically similar to the less well-documented Old Norse.)

For each persistence graph representing a language, the Betti number β_n gives the number of nth dimensional holes across all radii of the n-balls. Thus, β_0 represents the total number of clusters, and β_1 represents the total number of holes. [8] As n increases, β_n decreases; thus, there will be fewer holes than clusters for each language. Since not all IPA characters are used in a given language, the number of clusters and holes for different languages' persistence diagrams will vary.

In this study, the Bottleneck distance B is used to compare persistence diagrams:

$$\mathrm{B}_\infty(X, Y) = \inf_{\eta: X \to Y} \sup_{x \in X} ||x - \eta(x)|| \tag{6}$$

where X and Y are persistence diagrams showing a certain homology group and $\inf \sup$, or the infimum of the supremum, is the largest minimum distance between a point x in X and its bijection in Y, $\eta(x)$. [1]

Conceptually, the Bottleneck distance minimizes the maximum distance between two neighboring clusters or holes of two different diagrams, and B itself is that maximum distance. In the event that one language has more of one topological feature than the other, extra points are paired with the diagonal. [31]

A similar method of comparing persistence diagrams, sliced Wasserstein distance W, approximates distances between birth-death pairs by slicing them N number of ways and projecting them onto one-dimensional lines. Mathematically, it is defined as:

$$\mathrm{W}_\infty(X, Y) = \frac{1}{N} \sum_{i=1}^{N} ||\mu_i - \nu_i|| \tag{7}$$

where μ and ν are the distribution of points for X and Y, respectively, projected onto one-dimensional lines. [5] This metric is similar to the Bottleneck distance, but rather than taking

only the largest distance between two nearest-neighbors into account, it averages across all pairs of clusters or holes.

Validation

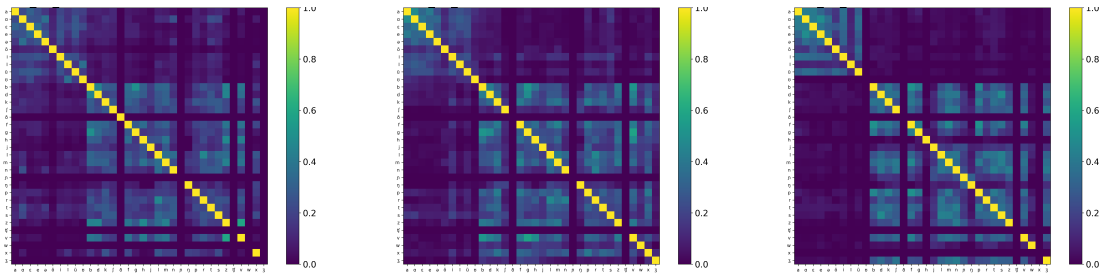

Figure 5: Visualization of cosine similarity matrices for German (left), Dutch (middle), and French (right) before TDA is performed on them.

To determine accuracy of the method detailed above, this study performed TDA on a set of thirteen modern Indo-European languages. Visually, the mistral-small translations for German and Dutch appear more similar through their cosine similarity matrices than German and French (see Figure 5). They can be compared quantitatively using the Frobenius norm metric F, which is defined as:

$$F = ||A - B||_F = \sum_{i=1}^{m} \sum_{j=1}^{n} |a_{ij} - b_{ij}|^2 \tag{8}$$

where A and B are cosine similarity matrices and (m,n) describes their dimensions. [9] Quantitatively, the Frobenius metric agrees with the visual colorplots, with F for German and Dutch being much smaller than F for German and French (see Figure 6).

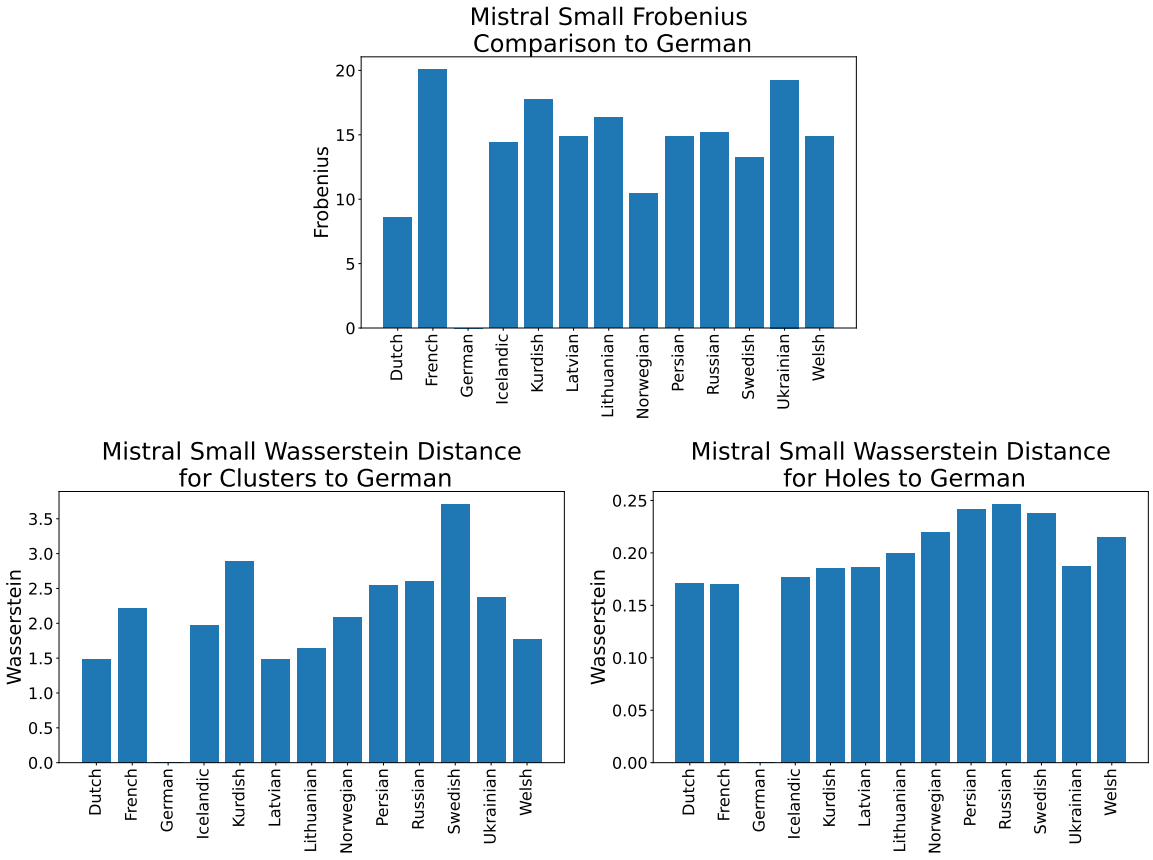

Figure 6: Frobenius norm between cosine similarity matrices compared to sliced Wasserstein distances after TDA.

Now we perform TDA on each cosine similarity matrix and compare persistence diagrams for the above languages using the sliced Wasserstein metric for clusters and holes (see Figure 6). While the more sensitive H_0 sliced Wasserstein distance agrees that German is more similar to Dutch than French, there are some deviations from the patterns in the Frobenius metric. However, this is to be expected, as the purpose of TDA is to detect higher-level information about the language's structure than other metrics (like the Frobenius metric) provide. Thus, validation confirmed that results make sense given general knowledge of modern languages, while also demonstrating the utility of TDA in analyzing data from a different perspective.

Because of biases introduced by variability in LLM translations and using one specific word list, the study must also investigate whether results can be distinguished from arbitrary results. If persistence diagrams do not differ greatly from randomly generated matrices, then results in this

study carry no meaning due to the arbitrary nature of obtaining data. Thus, this study follows Wolfram's (2017) approach in comparing TDA performed on randomly generated matrices with the same symmetry and diagonal as the input cosine similarity matrices. Figure 7 shows three such randomly generated matrices, and Figure 8 gives their corresponding persistence diagrams.

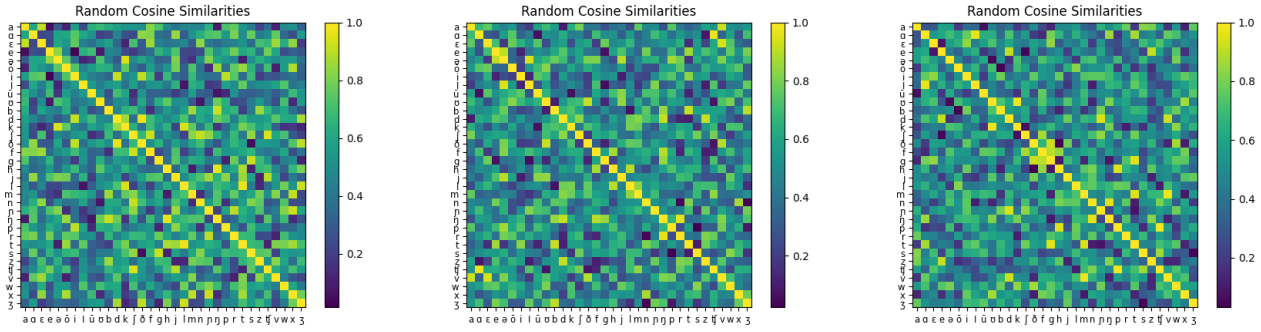

Figure 7: Random cosine similarity matrices generated in numpy and plotted in matplotlib. [11, 14]

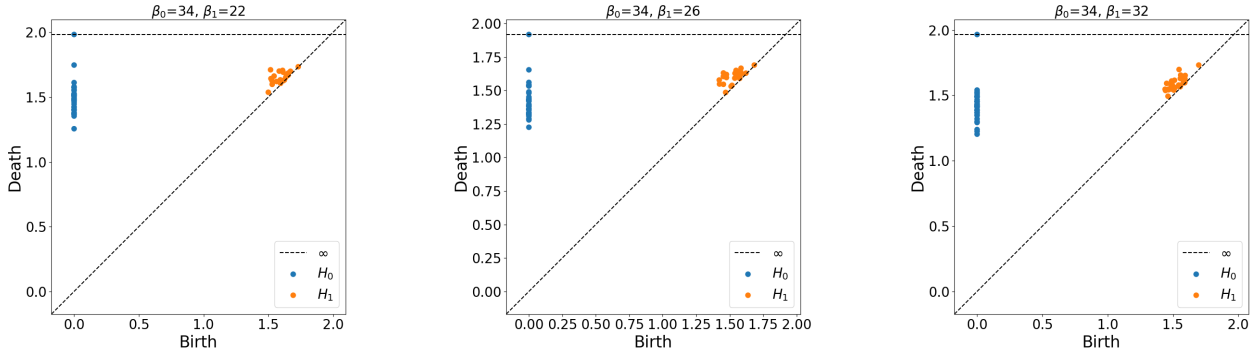

Figure 8: Persistence diagrams corresponding to the above cosine similarity matrices.

Persistence of clusters and holes in the random matrices did not match that of the language data; each topological feature appeared to have less variation in birth and death radii, and β_1 tended to be larger than for the language data. Thus, trends in my data were not random.

Results

Figure 9 gives the persistence diagram for Etruscan. Figure 10 then compares it using the bottleneck distance to each of the eight other languages across the six LLMs. While this is a valid and popular measure, this study found that it produced a very low distinguishability between data, especially for holes.

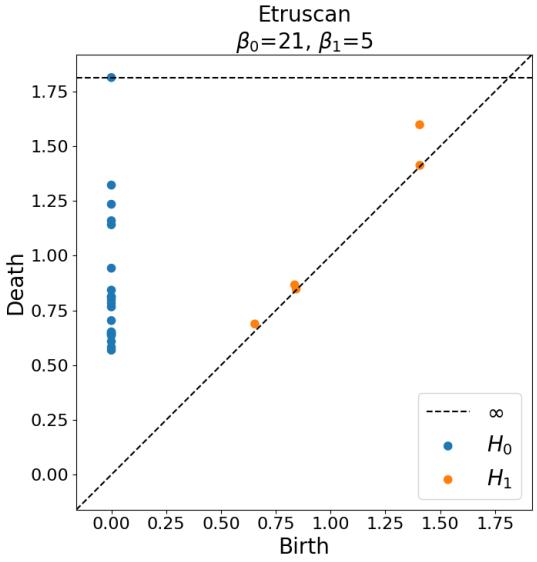

Figure 9: Persistence diagram for Etruscan.

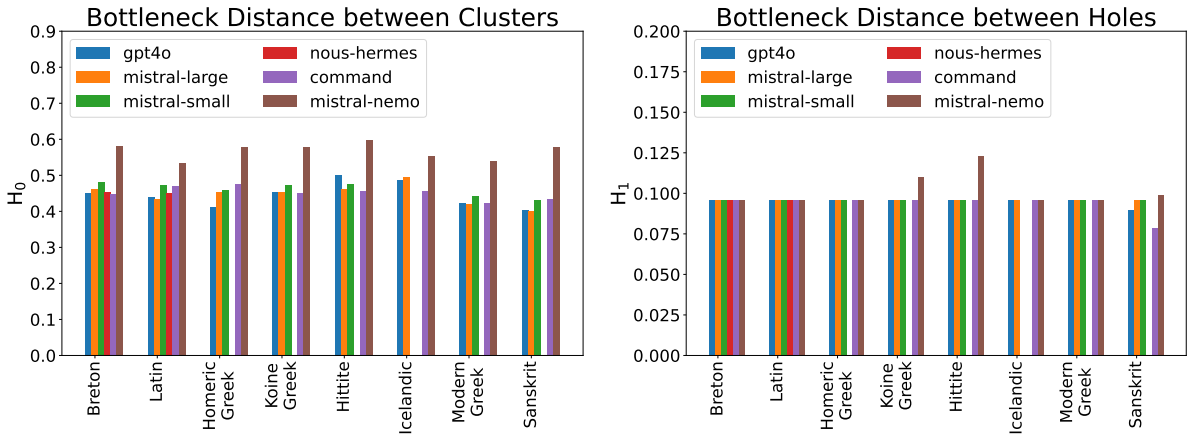

Figure 10: Bottleneck distances between Etruscan and eight other languages.

Figure 11 then performs the comparison using the more sensitive sliced Wasserstein distance.

While Figure 11 shows more variation in the data, there is still no clear trend.

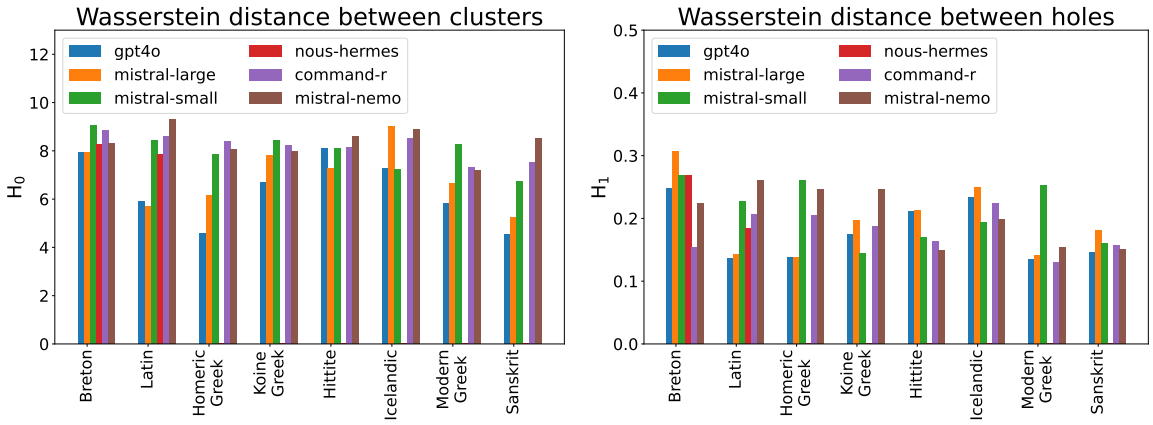

Figure 11: Sliced Wasserstein distances between Etruscan and eight other languages.

We must now consider the size of each LLM. To test the correlation between size of model and similarity in translations, each model's translation of the word list to Latin is compared to that of ChatGPT4o, the largest LLM. TDA is performed, and the sliced Wasserstein distances for H_0 and H_1 are shown in Figure 12. Mistral-large, the next largest model, produced the most similar Latin translations to that of ChatGPT4o. Thus, larger models will produce more similar translations. One can now hypothesize that as model size increases, a trend of the most similarity between Etruscan and a particular language from those tested will emerge.

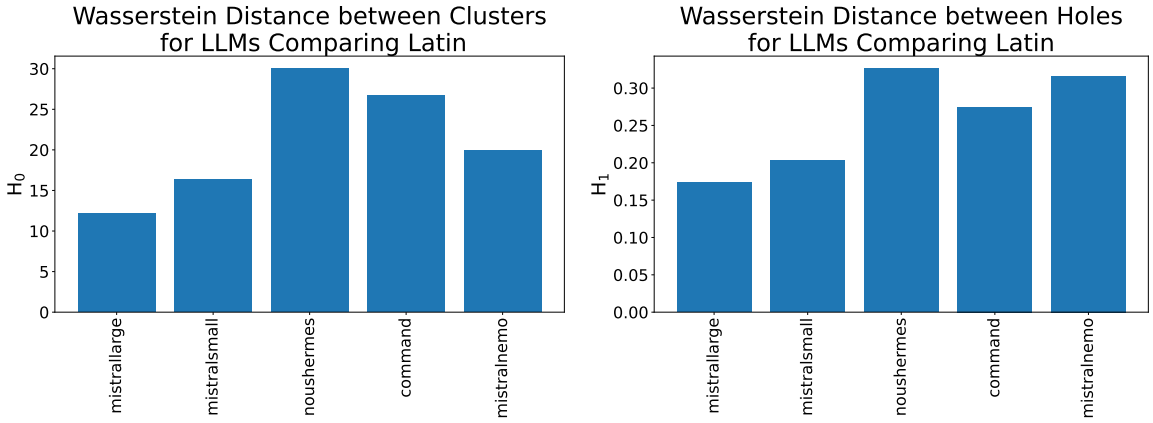

Figure 12: Comparison of LLMs to ChatGPT4o using Latin translations.

Figure 13 gives the sliced Wasserstein distance for H_0 between Etruscan and each of eight

other languages, arranged from greatest to smallest based on ChatGPT4o's output. As can be seen in the figure, ChatGPT4o translations point towards the Indo-Aryan language Sanskrit as being phonetically closest to Etruscan. A distinct trend can now be visualized; as models increase in size, they vary less from the results of ChatGPT4o, and the sliced Wasserstein distance between Sanskrit and Etruscan becomes progressively smaller (right side of Figure 13). Mistral-large and mistral-small agree that Sanskrit is closest to Etruscan. However, the smaller models of nous-hermes, command-r, and mistral-nemo show less of a trend, and other languages have a smaller sliced Wasserstein distance to Etruscan.

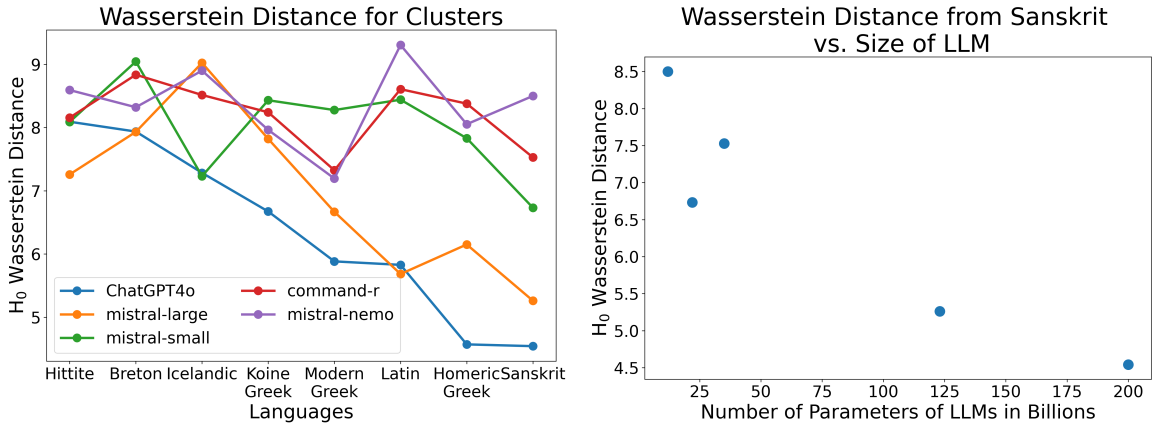

Figure 13: Sliced Wasserstein H_0 distance shows a correlation between Sanskrit and Etruscan.

The idea of Etruscan as a relative of Sanskrit is not currently a mainstream theory. There is little evidence that the Etruscans, supposed relatives of the Romans, were Aryans and spoke a variant of Sanskrit. How, then, do these results make sense? Since Sanskrit is the oldest Indo-European language considered here, it may be closer linguistically to the original Indo-European language, Proto-Indo-European (PIE), than other languages. Two other languages older than the majority of the dataset, Homeric Greek and Latin, appear next-most-similar to Etruscan using ChatGPT4o data. This is summarized in Figure 14, which shows a moderate correlation between age of the Indo-European language and its sliced Wasserstein distance to Etruscan. Thus, Etruscan appears most similar to the oldest of the Indo-European languages, indicating that it may be older than Latin and other languages spoken nearby. This would explain why the Etruscans' language appears unique compared to that of their genetically related neighbors.

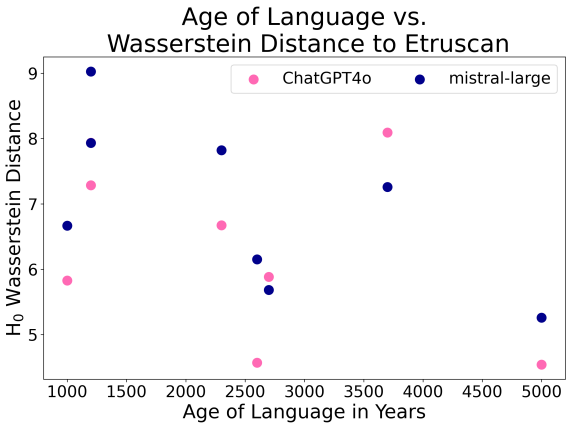

Figure 14: H_0 sliced Wasserstein distance vs. age of language compared to Etruscan.

Conclusions

Analysis of Results

In conclusion, results negate the Anatolian and Greek theories discussed in the Introduction, instead implying that Etruscan is most phonologically similar to older languages than to its neighbors. As such, it may have developed at an earlier time than some Indo-European languages. This would explain why the Etruscans lived in an Indo-European-speaking region and appear to be genetically related to the Romans but spoke a seemingly different language.

To test this hypothesis, access to phonological data for the original Indo-European language, PIE, is needed. However, PIE is a reconstructed language, and its IPA characters do not conform to those of better-understood languages. [17] In addition, LLMs even as large as ChatGPT4o may not provide translations to PIE to an adequately high percent of accuracy.

Analysis of Biases and Limitations

While this algorithm can be extended to a variety of other languages, the results obtained are dependent on the word list chosen, as well as the accuracy of the LLMs in providing translations. This study aimed to eliminate such biases by picking Etruscan words that are not proper names shared across languages, regardless of their inherent phonetic similarity, and utilizing multiple LLMs for translation to other languages. In addition, results could be skewed in the case of multiple IPA characters denoting the same phoneme, as the cosine similarity would not encode

all similarities that exist between two IPA characters. In fact, using more IPA characters than is linguistically accurate biases comparisons between persistence diagrams; the more noisy IPA characters are generated in a particular machine translation, the larger the bottleneck distance between that translation and Etruscan (see Figure 15).

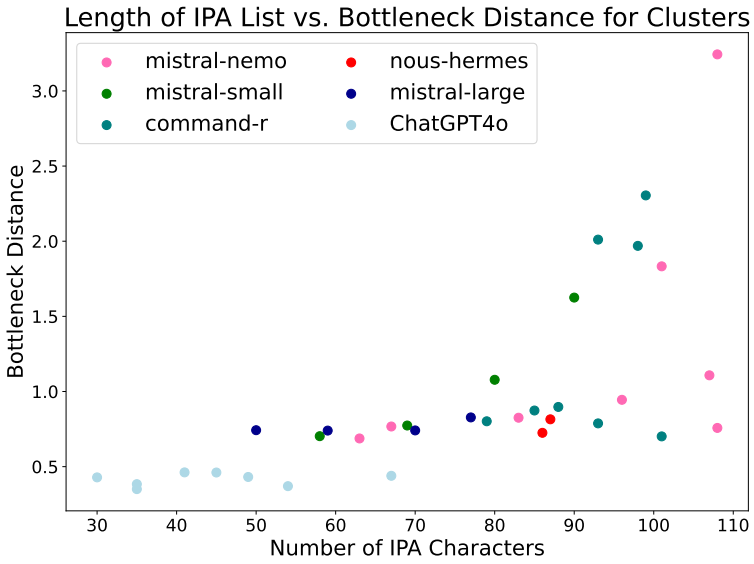

Figure 15: The greater the number of distinct IPA characters in a machine translation and language, the larger the Bottleneck distance between that translation and Etruscan.

To mitigate this bias, all IPA characters were researched, and superseded and nonstandard symbols were replaced to ensure that all translations of the word list conform to the same IPA encoding. Ultimately, this resulted in an IPA character list of 34 symbols.

Next Steps

Having constructed a working model for phonetic comparison between Etruscan and other languages, a more in-depth analysis of current findings and a study of Etruscan compared to a variety of other languages can be conducted. Immediate next steps also include comparing Etruscan to not only Indo-European languages but others language families (Semitic, Sino-Tibetan, etc.). In addition, more LLMs run on online servers, such as Claude and Gemini, will be used in collecting

translation data, and output of TDA will be analyzed through other metrics, such as the distribution of Betti numbers. [38] Hence accuracy of data will be improved.

Summary

Ultimately, the purpose of this project is threefold:

1. To determine the amount of variability between LLMs based on the parameter space in machine translations,
2. To discover the effectiveness of persistent homology as a more quantitative method for comparative historical linguistics, and
3. To conclusively compare the phonetic structure of Etruscan to various Indo-European languages in an effort to determine its origins and thus provide insight into the history behind the Etruscan people.

Appendix

Mathematical Glossary

topological data analysis (TDA): mathematical method of analyzing structure in data using topological features such as clusters and holes

persistent homology: form of TDA used in this study; analyzes how clusters and holes within a data structure persist over multiple scales

simplex: simplest method of connecting a given number of points in a given space; 0-simplex: point, 1-simplex: line segment, 2-simplex: triangle; etc.

n-chain: combination of simplices

birth radius: radius of n-balls drawn around each point in an nth-dimensional point cloud at which a topological feature is formed

death radius: radius of n-balls drawn around each point in an nth-dimensional point cloud at which a topological feature is consumed by another feature

persistence diagram: plot of death radius vs. birth radius for every cluster, hole, etc., that occurs in a point cloud

Linguistic Glossary

phonology: the system of sound values associated with different written characters in a given language

phoneme: a single sound associated with one or multiple written characters

Indo-European: language family from which the majority of Eurasian languages derive

Proto-Indo-European: a completely reconstructed language from which scholars believe all Indo-European languages derive

International Phonetic Alphabet (IPA): system for encoding phonemes as characters that are standardized across all languages

Acknowledgments

This project could not have been initiated without the helpful advice from my parents; their encouragement and the deeper understanding of science, computer science, and mathematics they have given me have been irreplaceable. I would also like to thank the many judges and reviewers from Science Fair and the Supercomputing Challenge, who generously provided me with feedback along the way, and the authors of the numerous papers that contributed to my knowledge of the project's background. Together they have given me the motivation to see this project to completion.

References

[1] Sarit Agami. Comparison of persistence diagrams, 2020. URL https://arxiv.org/abs/2003.01352.

[2] Simon Ager. Etruscan (mekh rasnal). https://www.omniglot.com/writing/etruscan.htm, Omniglot. 2023.

[3] Naomi Altman and Martin Krzywinski. The curse(s) of dimensionality. *Nature Methods*, 15 (6):399–400, 2018. doi: 10.1038/s41592-018-0019-x. URL https://doi.org/10.1038/s41592-018-0019-x.

[4] Giuliano Bonfante and Larissa Bonfante. *The Etruscan Language: an Introduction*. University of Manchester Press, 2002.

[5] Mathieu Carrière, Marco Cuturi, and Steve Oudot. Sliced wasserstein kernel for persistence diagrams, 2017. URL https://arxiv.org/abs/1706.03358.

[6] R Drikvandi and O Lawal. Sparse principal component analysis for natural language processing. *Ann. Data. Sci.*, 10(1):25–41, 2023. doi: 10.1007/s40745-020-00277-x.

[7] John Fraser. *The Etruscans: Were They Celts?* Maclachlan & Stewart, Edinburgh, 1879. URL https://archive.org/details/etruscanswerethe00fras.

[8] Ulderico Fugacci, Sara Scaramuccia, Federico Iuricich, and Leila De Floriani. Persistent homology: A step-by-step introduction for newcomers. 10 2016. doi: 10.2312/stag.20161358.

[9] Gene H. Golub and Charles F. Van Loan. *Matrix Computations*. Johns Hopkins, 3 edition, 1996.

[10] Aidan Gomez. Introducing command r7b: Fast and efficient generative ai. https://cohere.com/blog/command-r7b, 2024.

[11] Charles R. Harris, K. Jarrod Millman, Stéfan J. van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J. Smith, Robert Kern, Matti Picus, Stephan Hoyer, Marten H. van Kerkwijk, Matthew Brett, Allan Haldane, Jaime Fernández del Río, Mark Wiebe, Pearu Peterson, Pierre Gérard-Marchant, Kevin Sheppard, Tyler Reddy, Warren Weckesser, Hameer Abbasi, Christoph Gohlke, and Travis E. Oliphant. Array programming with NumPy. *Nature*, 585(7825):357–362, September 2020. doi: 10.1038/s41586-020-2649-2. URL `https://doi.org/10.1038/s41586-020-2649-2`.

[12] Csaba Horvath. Redefining pre-indo-european language families of bronze age western europe: A study based on the synthesis of scientific evidence from archaeology, historical linguistics and genetics. *European Scientific Journal*, 15(26):1–25, 2019. doi: 10.19044/esj.2019.v15n26p1.

[13] Caleb Howells. The surprising etruscan influence on the early celts. *The Collector*, 2023. URL `https://www.thecollector.com/etruscan-influence-celts-connection/`. accessed Jan 24, 2025.

[14] J. D. Hunter. Matplotlib: A 2d graphics environment. *Computing in Science & Engineering*, 9(3):90–95, 2007. doi: 10.1109/MCSE.2007.55.

[15] Milorad Ivankovic. Etruscan as an aryan indo-european language. 11 2021.

[16] Eamonn Keogh and Abdullah Mueen. Curse of dimensionality. *Enclyclopedia of Machine Learning and Data Mining*, 2017.

[17] Winfred P. Lehmann. Proto-indo-european phonology: 2. pie phonology. https://lrc.la.utexas.edu/books/piep/2-pie-phonology, 2005.

[18] maix. Image own work based on: Europe countries.svg by tintazul, cc by-sa 3.0. https://commons.wikimedia.org/w/index.php?curid=1636794.

[19] Rick McCallister and Silvia McCallister-Castillo. Etruscan glossary, 1999. URL https://etruscans1.tripod.com/Language/EtruscanIntro.html. accessed Jan 24, 2025.

[20] Mistral AI. Au large. https://mistral.ai/news/mistral-large, 2024.

[21] Mistral AI. Mistral nemo. https://mistral.ai/news/mistral-nemo, 2024.

[22] Mistral AI. Mistral small 3. https://mistral.ai/news/mistral-small-3, 2025.

[23] Prerna Nadathur. An introduction to homology. 2007. URL https://api.semanticscholar.org/CorpusID:1431656.

[24] OLLAMA Website. Ollama. https://ollama.com/, 2025.

[25] OpenAI. Chatgpt4o (may 24 version). https://platform.openai.com/docs/overview, 2022.

[26] Alexander Port, Iulia Gheorghita, Daniel Guth, John M. Clark, Crystal Liang, Shival Dasu, and Matilde Marcolli. Persistent topology of syntax, 2015. URL https://arxiv.org/abs/1507.05134.

[27] Alexander Port, Iulia Gheorghita, Daniel Guth, John M. Clark, Crystal Liang, Shival Dasu, and Matilde Marcolli. Persistent topology of syntax. *Mathematics in Computer Science*, 12 (1):33–50, 2018. doi: 10.1007/s11786-017-0329-x. URL https://doi.org/10.1007/s11786-017-0329-x.

[28] Cosimo Posth, Valentina Zaro, Maria A. Spyrou, Stefania Vai, Guido A. Gnecchi-Ruscone, Alessandra Modi, Alexander Peltzer, Angela Mötsch, Kathrin Nägele, Åshild J. Vågene, Elizabeth A. Nelson, Rita Radzevičiūtė, Cäcilia Freund, Lorenzo M. Bondioli, Luca Cappuccini, Hannah Frenzel, Elsa Pacciani, Francesco Boschin, Giulia Capecchi, Ivan Martini, Adriana Moroni, Stefano Ricci, Alessandra Sperduti, Maria Angela Turchetti, Alessandro Riga, Monica Zavattaro, Andrea Zifferero, Henrike O. Heyne, Eva Fernández-Domínguez, Guus J. Kroonen, Michael McCormick, Wolfgang Haak, Martina

Lari, Guido Barbujani, Luca Bondioli, Kirsten I. Bos, David Caramelli, and Johannes Krause. The origin and legacy of the etruscans through a 2000-year archeogenomic time transect. *Science Advances*, 7(39):eabi7673, 2021. doi: 10.1126/sciadv.abi7673. URL https://www.science.org/doi/abs/10.1126/sciadv.abi7673.

[29] Helmut Rix. *Etruscan*, page 141–164. Cambridge University Press, 2008.

[30] Adelle Rogers. Theories on the origin of the etruscan language. *Open Access Theses*, page 1587, 2018. URL https://docs.lib.purdue.edu/open_access_theses/1587.

[31] Nathaniel Saul, Michael Catanzaro, Kostya Lyman, and Roberto Panai. scikit-tda/scikit-tda: v1.1.1, jul 2024.

[32] Ulrich Schädler. Greeks, etruscans, and celts at play. *Archimède. Archéologie et histoireancienne*, 6:160–174, 2019. doi: 10.47245/archimede.0006.ds2.08.

[33] Rohit P. Singh, Nicholas O. Malott, Blake Sauerwein, Neil Mcgrogan, and Philip A. Wilsey. Generating high dimensional test data for topological data analysis. In Sascha Hunold, Biwei Xie, and Kai Shu, editors, *Benchmarking, Measuring, and Optimizing*, pages 18–37, Singapore, 2024. Springer Nature Singapore.

[34] Shaun Sutner. Openai advances llm with gpt-4o; google gemini update looms. https://www.techtarget.com/searchenterpriseai/news/366584023/OpenAI-advances-LLM-with-GPT-4o-Google-Gemini-update-looms, 05 2024.

[35] Funmilola Mary Taiwo, Umar Islambekov, and Cuneyt Gurcan Akcora. Explaining the power of topological data analysis in graph machine learning, 2024. URL https://arxiv.org/abs/2401.04250.

[36] Ryan Teknium, Jeffrey Quesnelle, and Guang Chen. Hermes 3 technical report. https://arxiv.org/abs/2408.11857, 2024.

[37] Bouke Van der Meer. Etruscan origins: Language and archaeology. *BABESCH*, 79:51–57, 2004. doi: 10.2143/BAB.79.0.504735.

[38] Catherine Wolfram. Persistent homology on phonological data: a preliminary study. https://math.uchicago.edu/ may/REU2017/REUPapers/Wolfram.pdf, 2017.