Interspeech 2024
1-5 September 2024, Kos, Greece

Searching for Structure: Appraising the Organisation of Speech Features in wav2vec 2.0 Embeddings

Patrick Cormac English[1,2], *John D. Kelleher*[3], *Julie Carson-Berndsen*[2]

[1]SFI Centre for Research Training in Digitally-Enhanced Reality (d-real), Ireland, [2]ADAPT Research Centre, School of Computer Science, University College Dublin, Ireland, [3]ADAPT Research Centre, School of Computer Science and Statistics, Trinity College Dublin, Ireland

patrick.english@ucdconnect.ie, julie.berndsen@ucd.ie, John.Kelleher@tcd.ie

Abstract

Recent advancements in speech recognition have been driven by large transformer models trained on extensive unlabelled speech corpora. These models generate speech representations that potentially encapsulate key speech features, yet the organisation of these features within the model's embedding space and their alignment with phonetic and phonological theories remains unclear. This paper aims to bridge this gap by applying probing methods to explore the structure of phonetic information within embeddings, thereby uncovering linguistically significant relationships within the latent representations. We introduce a novel approach that probes the speech embeddings for independent features and then applies association rule mining to identify relationships and organisational structure within the data. Our research seeks to enhance the understanding of the speech embeddings of transformer models, ultimately contributing to the explainability of these systems.

Index Terms: transformer-based speech recognition, probing, phonetic and phonological features

1. Introduction

Recent advancements in self-supervised learning, and specifically transformer models leveraged for this task, have led to the development of promising speech representations, or embeddings, such as those generated by wav2vec 2.0. It has been shown that these embeddings encode rich linguistic information, including phonetic details, without explicit supervision. Building upon other work that has explored the relationship between large transformer embeddings and phonetic information through probing tasks, this paper aims to delve deeper into the phonetic and phonological knowledge encoded within wav2vec 2.0 embeddings.

This paper investigates whether the feature matrices proposed by phonologists and the relationships that were identified in phonological and phonetic theories, are indeed encoded in modern transformer-based speech recognition models. Our approach involves the development of independently trained probing models, designed to identify the presence of phonological organisation of features within wav2vec 2.0 embeddings. Previous research has demonstrated that individual time-steps in wav2vec 2.0 embeddings encode sub-phonetic information i.e. below the level of the phone; thus, we aim to investigate whether this information is structured in a way that aligns with theoretical expectations derived from phonological frameworks. By training probes on a set of features for which we have a well-established phonological framework, we propose an extension to existing probing methods (see Figure 1) that enables the investigation of whether the patterns of probe activations for these (sub-phonetic) features within our data align with theory-informed expectations. To this end, we employ association rule mining to uncover potential co-occurrence rules and patterns within the probe activations, and compare them to rules similarly mined from our framework itself.

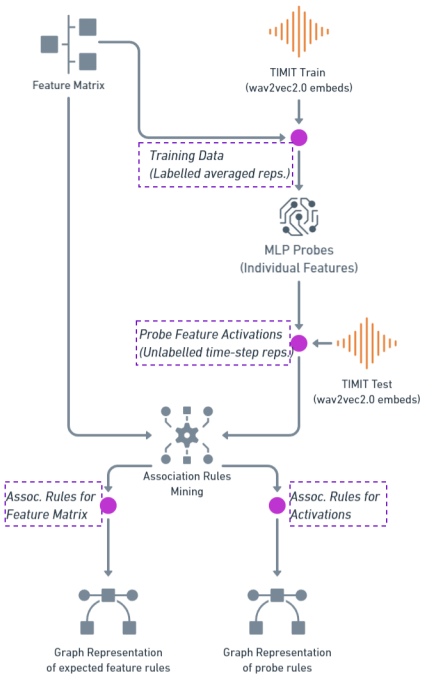

Figure 1: *Visualisation of the Experimental Methodology*

The paper is structured as follows. Section 2 presents related work including the feature matrix we use in our experiments. Section 3 describes the TIMIT data and the training procedure for the probes used in our experiments. Section 4 introduces association rule mining and discusses its application to our feature matrix and probe activations. Finally, Section 5 describes the graph representation process and the resulting visualisations.

2. Related Work

There are two main areas of related work. The first is the emergence of domain-informed probing as a means of interrogating the pre-trained embeddings of larger transformer-based models, in natural language processing [1, 2] and now more recently in speech technology [3, 4, 5, 6, 7]. One of the key motivations for probing speech transformers is to identify whether these models have encoded phonetic and phonological knowledge dur-

10.21437/Interspeech.2024-2047

Table 1: *Subsection of Feature Matrix based on the IPA*

	p	t	b	d	k	g	f	v	s	sh
vowel	-	-	-	-	-	-	-	-	-	-
consonant	+	+	+	+	+	+	+	+	+	+
plosive	+	+	+	+	+	+	-	-	-	-
nasal	-	-	-	-	-	-	-	-	-	-
fricative	-	-	-	-	-	-	+	+	+	+
bilabial	+	-	+	-	-	-	-	-	-	-
labiodental	-	-	-	-	-	-	+	+	-	-
dental	-	-	-	-	-	-	-	-	-	-
alveolar	-	+	-	+	-	-	-	-	+	-
postalveolar	-	-	-	-	-	-	-	-	-	+
palatal	-	-	-	-	-	-	-	-	-	-
velar	-	-	-	-	+	+	-	-	-	-
voicing	-	-	+	+	-	+	-	+	-	-

ing training. Moving beyond this existing work, which examined whether transformer embeddings encoded phonetic information, we examine whether there is a (theoretically aligned) structure to how phonetic features[1] are organised in the speech embeddings, in particular, whether the pattern of encoding of these features align with the constraints that arise from phonological theory. For this reason much of this section is devoted to the second area of related work, namely those aspects of phonetic and phonological theory which are relevant to the approach we present in Section 4. There has long been a discussion in phonology regarding which features are most appropriate for describing the phonological processes which can be found in languages of the world. Each language is assumed to have a specific set of features which describe the phonemes of that language. We are not providing input as to what may be considered to be the best feature set. However, we note that, in general, features can be classified as unary, binary and multi-valued features. Unary features are descriptors without any structure and are thus difficult to relate to each other without making further assumptions. Binary features denote presence vs. absence of properties, although again there has been a debate on whether absence of a property can really constitute a characteristic. Multi-value features are those which demonstrate the most structure in that they capture attributes with values and can be found primarily in non-segmental approaches to phonology e.g. in *autosegmental phonology* [8] where the attributes are tiers and the values are properties on those tiers, or in *articulatory phonology* [9] where the attributes are articulators and the values describe what the specific articulator is doing. The IPA classification of sounds [10] can be interpreted as multi-valued features, with the categories of *manner of articulation* (MOA), *place of articulation* (POA) and *voicing* as attributes with specific values for each of the categories, such as plosive or fricative for MOA.

One of the main advantages of using features in phonetics and phonology is that it is possible to form related classes which pattern together; these are known as natural classes in phonology or broad phonetic groups in phonetics. The mapping between phonemes of a given language and chosen features can be represented in a tabular structure known as a *Feature Matrix*. In segmental phonology, based on these matrices, it was possible to construct classes of features which represent more than one phoneme and define co-occurrence relationships between the phonemes in what were termed redundancy rules [11]. In non-segmental phonology, the relationships between the values were described in terms of associations [12] and temporal events [13].

[1]Note that we use the term phonetic here since the features are coming from the speech.

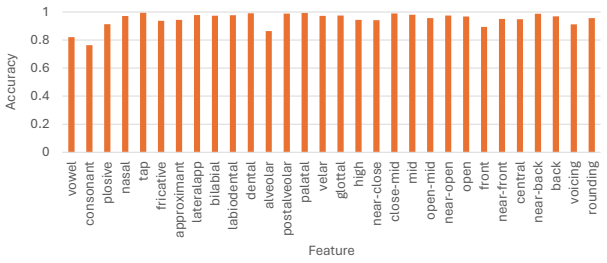

Figure 2: *Probe Accuracies per Feature for layer 9*

This paper does not focus on a specific theory, but instead looks to decipher the organisation of phonetic features and relationships among them through the use of probing. Rather than making the assumption that the features for the probes are grouped together as in the IPA classification into MOA, POA and voicing ([6, 7]) where probes were trained independently for each of these classes, our approach flattens the IPA classification to a set of binary features, i.e. rather than *MOA: plosive* and *POA: bilabial*, we have + *plosive*, - *plosive*, + *bilabial*, - *bilabial* (see Table 1). We believe that by using a flat binary classification, we can better identify whether the pre-trained model has really learned relationships between the features without being provided structure as regards how these relate to MOA, POA and voicing. In what follows, we distinguish between a static distinctive feature matrix, which defines the mapping between phonemes of English and features based on the IPA classification, and a set of independent probes which are trained on TIMIT to recognise these features. The feature matrix represents the knowledge-based classification of the features by the phonologist, and the probes look to see whether the feature organisation is encoded in the transformer embeddings.

3. Probe Training

The following section details the data and labelling design for input to the probes used to generate activations for features. It also outlines the configuration of the probe models.

3.1. Data & Embeddings

The TIMIT read-speech corpus [14] was used for this work, containing 5.4 hours of 16 kHz spoken American-accented English audio in wav format. The entire TIMIT dataset was used, following the TIMIT suggested training/test split. Wav2vec 2.0 [15] was employed to extract embeddings from the TIMIT audio data. The embeddings were obtained from layer 9 of the wav2vec 2.0 model and averaged over the duration of each phone, resulting in phone-level representations. This layer of wav2vec 2.0 has been identified in our prior work examining phonetic information capture as having high performance for tasks of the nature undertaken in this work [6].

3.2. Probe Architecture

30 multilayer perceptron (MLP) models were trained to predict the presence of 30 distinct phonetic features based on the IPA schema from the phone-averaged wav2vec 2.0 embeddings. The scikit-learn [16] implementation of the MLP was used, with 1 hidden layer of 200 ReLU activation neurons and a sin-

gle output neuron with logistic activation function. Each MLP was trained on wav2vec 2.0 25ms (20ms stride) representations (a [1*768] vector), averaged by individual phone following the methodology implemented in [3, 6], and provided with the feature-presence label as a supervision signal. The training dataset consisted of 175,232 phone-averaged samples from the TIMIT training set, and the 258,040 samples from the TIMIT test set were reserved for evaluation. The sample embeddings were labelled during training based on a mapping of the existing TIMIT phone label (from the averaging process) to the training feature. Figure 2 plots the accuracy of the trained probes on the individual time-steps, which were labelled based on the TIMIT metadata of the phone occurring during that time-step. To ensure probe performance reflected task-relevant information, a separate set of randomly generated datasets was created [17]. The probes trained on these randomised datasets were unable to effectively predict features and performed close to the random chance baseline.

4. Association Rule Mining

Association rule mining [18], common in the analysis of large-scale market transactions, is a simple machine learning technique for identifying dependencies between features. This was implemented via the `Arulespy` package `Arules` implementation [19]. The algorithm used for rule identification was `Apriori` [18]. We apply this technique at two different points.

Feature Matrix: Firstly, given the knowledge-based distinctive feature matrix (FM) which defines the mapping between English phonemes and features, we apply association rule mining to see what co-occurrence relationships exist between the phonological features. This is similar to the construction of the redundancy rules or feature implications of phonology. We consider this mapping the ground truth for the relationships between features

Probe Activation: Secondly, we apply the same technique to the output of the probes on data-driven embeddings (see Section 4.2) to determine whether similar patterns are found between the phonetic features on an English dataset, namely TIMIT.

It is important to note the distinction here between the FM which describes each phoneme in terms of properties in the IPA classification and the probes which have been trained independently of each other on TIMIT.

4.1. Association Rule Mining of the Feature Matrix

Table 2 shows the top outputs for the association rule mining showing feature co-ocurrences in the FM. All but one of these have a confidence of 1. The first rule shows an association between + *vowel* and + *voicing* and is interpreted as an implication:

$$vowel \Rightarrow voicing$$

which would be expected, since all vowels are indeed voiced. More interesting perhaps, is the rule

$$alveolar \Rightarrow voicing$$

with a confidence of 0.81. This is a result of an imbalance of the number of phonemes defined in the FM with the features + *alveolar* and + *voicing* (9 phonemes) as opposed to + *alveolar* and - *voicing* (only 2 phonemes). Support indicates the number of phonemes/phones which are impacted by an association rule, and lift refers to the strength of the relationship between the features. A lift value greater than 1 indicates a positive association between the features.

Table 2: *Top association rules learned from the FM ranked by confidence. Label indicates + feature; LHS and RHS refer to left-hand side and right-hand side of the rule respectively, C is confidence, S is support and L is lift*

Count	LHS	RHS	C	S	L
20	vowel	voicing	1	0.36	1.31
12	fricative	consonant	1	0.22	1.72
11	alveolar	consonant	1	0.20	1.72
7	plosive	consonant	1	0.13	1.72
7	nasal	consonant	1	0.13	1.72
7	nasal	voicing	1	0.13	1.31
7	front	vowel	1	0.13	2.75
9	alveolar	voicing	0.81	0.16	1.07

Table 3: *Top association rules learned from the probe activations, ranked by confidence*

Count	LHS	RHS	C	S	L
68822	vowel	voicing	0.99	0.27	1.5
25790	front	voicing	0.99	0.1	1.51
21348	approximant	voicing	0.98	0.08	1.5
33556	fricative	consonant	0.96	0.13	1.67
38655	plosive	consonant	0.95	0.15	1.66
19934	approximant	consonant	0.92	0.08	1.6
74194	alveolar	consonant	0.90	0.29	1.57
22644	front	vowel	0.87	0.09	3.22

4.2. Association Rule Mining of the Probe Activations

The trained probes were applied to the time-step level TIMIT test representations (see Section3.2), predicting the presence or absence of phonetic features at each time step. This generated, for each time-step representation, a vector of dimension [1 * 30], corresponding to a [+ -] value for each feature probe. A combined vector of [258,040 * 30], 258,040 being the number of time-steps evaluated, was then processed via the methods described in the previous section to generate rules for the test data.

Table 3 shows the top outputs for the association rule mining on the feature probe activations. Many rules are shared in the top-count rules, with the top-4 rules for the FM present in the top probe-activation rules.

5. Visualisation of Feature Organisation

To visualise the discovered association rules, we built a graph representation using the `ArulesViz` [19] package. In this graph, each node represents a phonetic feature, and the edges represent the association rules connecting the features. The edges are unweighted and indicate the presence of a rule between two features. To focus on the most relevant rules, we filtered the graph to include only rules with a single antecedent. This means that each edge in the graph represents a rule of the form "if feature A is present, then feature B is likely to be present", i.e.

$$+A \Rightarrow +B$$

The graph visualisation also incorporates two of the metrics identified earlier: lift and support. As noted in Section 4.1, the lift of a rule measures the ratio of the observed co-occurrence of the antecedent [LHS] and consequent [RHS] features to the expected co-occurrence if they were independent. The support

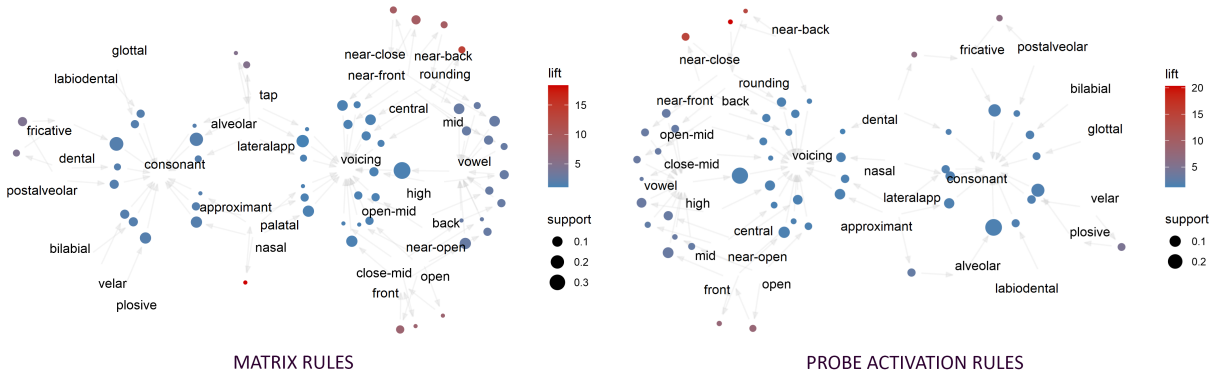

Figure 3: *Graph Visualisation of Rules*

of a rule represents the proportion of instances in the dataset where both the antecedent and consequent features are present together. The lift and support values are displayed on the rule nodes of the graph, providing insights into the strength and significance of each association rule. Lift is plotted as a blue-red (low-high) gradient on rule nodes, with node size representing support. The graph visualisation provides a concise, yet interpretable, representation of the discovered association rules, highlighting the most prominent and meaningful relationships between the phonetic features. It enables a quick identification of strongly associated features and facilitates the comparison of the rules obtained from the ground truth mapping (FM) and the predicted feature presence.

5.1. Analysis of Associations

The graphical representations of the association rules identified in the feature matrix (FM) and probe activations allow for a high-level analysis of the phonological structure captured by wav2vec 2.0 embeddings. By visualising the association rules as graphs, we can identify patterns such as paths, clusters, and centrality of features, which may be less apparent in the tabular format presented in Section 4. It also allows for ready comparison of the two rule-sets at a high-level. Figure 3 illustrates these graphical representations, with the relative position of feature nodes and links reflecting the information within the mined rules. Clusters indicate the convergence of multiple association rules onto feature nodes, with a node receiving multiple edges indicative of a consequent feature which is the consequent of many separate antecedents in the rule-set. The significant similarity between the graph structures generated for the two rule-sets extends beyond the top rules discussed earlier. Both visualisations display comparable clustering of associations, indicating that the similarities observed in the most significant rules are maintained throughout the entire structure. This consistency suggests that wav2vec 2.0 embeddings capture a robust and coherent phonological structure.

In the MATRIX RULES visualisation, the consonant and vowel features form distinct clusters, with the voicing feature serving as a central alignment point. This structure highlights the significant overlap between the voicing feature and the otherwise separated consonant and vowel feature clusters. The well-defined rules in this representation indicate strong and unambiguous relationships between constituents of the FM. The probe activation visualisation exhibits a similar general structure, with some difference in the organization of peripheral features. Lift values remain similar for equal rules in each rule-set, indicating that the strengths of co-occurence relationships in the FM rule-set are similarly identified in the probe activations.

Despite the more complex and less delineated organisation for certain features in the PROBE ACTIVATION RULES visualisation, which reflects the challenges in capturing fine-grained phonological distinctions through probing models, it is clear that the overall structure of the associations found in both rule-sets remains largely congruent. These similarities observed in both the rule-sets and the graphical representations underscore the robustness of the phonological structure captured by wav2vec 2.0 embeddings. Notably, of the 66 rules in our Feature Matrix rule-set, 49 were present within the rule-set mined from the probe activations.

6. Conclusion and Future Work

The findings of this study demonstrate the potential of using association rules and co-activation patterns to explore the phonological knowledge captured within wav2vec 2.0 embeddings. One promising avenue for future research is to investigate the value of implications between features for inheritance. By examining the co-occurrence and dependency relationships among phonological features, we can gain insights into how certain features are inherited or derived from others. This approach could better explore on the presence of hierarchical structure within embedding representations of large self-supervised models. Future work would also seek to expand the processes described here to a wider group of candidate model architectures, in addition to exploring the encoding of association information across layers in these transformer architectures. It would also seek to perform an analysis on less-expected outputs in the rules mined from probe output, to supplement the high-level analysis performed here.

We also note the limitations of the probing methodology, as probe accuracy and training design introduce bounds on the granularity of information that may be identified within the embedding space. Future work will seek to explore methods for improved operation of these probes with respect to our tasks. Moreover, the current study focused on a specific set of phonological features and their associated theoretical framework. To further validate and expand upon our findings, future work should explore other feature sets and geometries to assess the generalisability of our approach and uncover potential variations in the way embeddings capture phonological information across diverse feature spaces.

7. Acknowledgements

This work was conducted with the financial support of Science Foundation Ireland at the SFI Centre for Research Training in Digitally-Enhanced Reality (d-real) [18/CRT/6224], and at ADAPT, the SFI Research Centre for AI-Driven Digital Content Technology at University College Dublin [13/RC/2106_P2]. For the purpose of Open Access, the author has applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission.

8. References

[1] F. Klubička, V. Nedumpozhimana, and J. D. Kelleher, "Idioms, Probing and Dangerous Things: Towards Structural Probing for Idiomaticity in Vector Space," *arXiv:2304.14333v1*, 2023. [Online]. Available: https://arxiv.org/abs/2304.14333v1

[2] T. Xiao and J. Zhu, "Introduction to transformers: an nlp perspective," 2023.

[3] P. C. English, J. D. Kelleher, and J. Carson-Berndsen, "Domain-Informed Probing of wav2vec 2.0 Embeddings for Phonetic Features," in *SIGMORPHON*, 2022, pp. 83–91. [Online]. Available: https://aclanthology.org/2022.sigmorphon-1.9

[4] M. Yang, R. C. M. C. Shekar, O. Kang, and J. H. L. Hansen, "What Can an Accent Identifier Learn? Probing Phonetic and Prosodic Information in a Wav2vec2-based Accent Identification Model," *arXiv:2306.06524v1*, 2023. [Online]. Available: http://arxiv.org/abs/2306.06524

[5] L. ten Bosch, M. Bentum, and L. Boves, "Phonemic competition in end-to-end ASR models," in *INTERSPEECH*, 2023, pp. 586–590. [Online]. Available: https://www.isca-speech.org/archive/interspeech_2023/tenbosch23_interspeech.html

[6] P. C. English, J. D. Kelleher, and J. Carson-Berndsen, "Discovering Phonetic Feature Event Patterns in Transformer Embeddings," in *INTERSPEECH*, 2023.

[7] P. C. English, E. A. Shams, J. D. Kelleher, and J. Carson-Berndsen, "Following the embedding: Identifying transition phenomena in wav2vec 2.0 representations of speech audio," in *ICASSP*, 2024.

[8] J. A. Goldsmith, "Autosegmental phonology," Ph.D. dissertation, Massachusetts Institute of Technology, 1976.

[9] C. P. Browman and L. Goldstein, *Tiers in articulatory phonology, with some implications for casual speech*, ser. Papers in Laboratory Phonology. Cambridge University Press, 1990, p. 341–376.

[10] "IPA Chart, available under a Creative Commons Attribution-Sharealike 3.0 Unported License. Copyright © 2015 International Phonetic Association." [Online]. Available: https://www.internationalphoneticassociation.org/content/full-ipa-chart

[11] R. Stanley, "Redundancy rules in phonology," *Language*, vol. 43, no. 2, pp. 393–436, 1967. [Online]. Available: http://www.jstor.org/stable/411542

[12] E. C. Sagey, "The representation of features and relations in non-linear phonology," Ph.D. dissertation, Massachusetts Institute of Technology, 1986.

[13] S. Bird and E. Klein, "Phonological events1," *Journal of linguistics*, vol. 26, no. 1, pp. 33–56, 1990.

[14] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, N. L. Dahlgren, and V. Zue, "TIMIT acoustic-phonetic continuous speech corpus," *Linguistic Data Consortium*, 1993. [Online]. Available: https://catalog.ldc.upenn.edu/LDC93S1

[15] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in *NeurIPS*, vol. 33, 2020, pp. 12 449–12 460. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2020/file/92d1e1eb1cd6f9fba3227870bb6d7f07-Paper.pdf

[16] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[17] Y. Belinkov and J. Glass, "Analyzing hidden representations in end-to-end automatic speech recognition systems," in *NeurIPS*, vol. 30, 2017. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2017/file/b069b3415151fa7217e870017374de7c-Paper.pdf

[18] R. Agrawal, T. Imieliński, and A. Swami, "Mining association rules between sets of items in large databases," *SIGMOD Rec.*, vol. 22, no. 2, p. 207–216, jun 1993. [Online]. Available: https://doi.org/10.1145/170036.170072

[19] M. Hahsler, "Arulespy: Exploring association rules and frequent itemsets in python," 2023.