Analysis and comprehensive lineage identification for SARS-CoV-2 genomes through scalable learning methods

Roberto Cahuantzi[a,1], Katrina Lythgoe[b, c], Ian Hall[a,d], Lorenzo Pellis[a,d], and Thomas House[a,d]

[a]Department of Mathematics, University of Manchester, Manchester, UK.
[b]Department of Biology, University of Oxford, Oxford, UK.
[c]Big Data Institute, University of Oxford, Oxford, UK.
[d]The Alan Turing Institute, London, UK.
[1]Corresponding Author: roberto.cahuantzi@manchester.ac.uk

Abstract

Since its emergence in late 2019, SARS-CoV-2 has diversified into a large number of lineages and globally caused multiple waves of infection. Lineages have the potential to spread rapidly and internationally if they have higher intrinsic transmissibility and/or can evade host immune responses, as has been seen with the Alpha, Delta, and Omicron variants of concern (VoCs). Phylogenetic methods provide the gold standard for representing the global diversity of SARS-CoV-2 and to identify newly emerging lineages. However, these methods are computationally expensive, struggle when datasets get too large, and require manual curation to designate new lineages. These issues will only intensify as the vast number of SARS-CoV-2 genomes already available continues to grow. It will therefore be beneficial to develop complementary methods that can incorporate all of the genetic data available, without down sampling, to extract meaningful information rapidly and with minimal curation. Here, we demonstrate the utility of using algorithmic approaches based on word-statistics to represent whole sequences, bringing speed, scalability, and interpretability to the construction of genetic topologies, and that can be used to augment traditional classification practice based on phylogeny.

Introduction

The rapid spread of SARS-CoV-2 during 2020 resulted in major healthcare and societal challenges at the global level. More recently, the periodic emergence of new variants that are more transmissible and/or escape host immune responses has given rise to repeated waves of infection that have each produced considerable burdens of disease despite high rates of vaccination and prior infection history [33]. There is now extensive effort in identifying worrying new variants at the very earliest stages of their emergence, which at best may enable elimination of these variants before they become established [31], but otherwise enables forward planning that may, in the future, include the timely production of tailored vaccines.

SARS-CoV-2, like other RNA viruses, has a high mutation rate and a short generation time, meaning it evolves extremely rapidly and on the same timescale as transmission. As a consequence, phylogenetics has been a powerful approach to monitor the evolution of SARS-CoV-2 [27, 24, 7]. Most mutations, or single-nucleotide polymorphisms (SNPs), that appear on phylogenetic trees are neutral or nearly neutral, meaning the mutations themselves have little or no impact on transmission, which lineages and their defining SNPs are successful and which die out are defined by selection pressures. However, some mutations do have a selective advantage because they enable more effective transmission in the population at the time of emergence. A hallmark of SARS-CoV-2 evolution has been the emergence and spread of variants of concern (VoCs) and some of their major sublineages, with each having a large collection, or constellation, of lineage defining mutations, many of which have given these viral

lineages transmission advantages that have enabled them to spread extremely effectively [3].

Identifying viral lineages that are likely to be problematic in the future requires considerable effort [27], with sequences requiring alignment to a reference before incorporating into a phylogenic topology, designating (naming) new lineages, and identifying lineages with potentially troublesome mutations or that are expanding quickly [7]. The production of phylogenies containing all high-quality SARS-CoV-2 genomes could help in the automation of this process, but with over 12 million sequences available in the GISAID database [30], as of September 2022, and growing, aligning all of the sequences and generating a single phylogeny is a near impossible task. Alignment-free methods potentially offer a computationally much faster option for designating new lineages and that can cope with massive amounts of genomic data [10, 41, 40, 34, 38, 8, 13, 2].

Alignment-free techniques to characterise sequences can be classified into two types: information-theory and word-statistics based. Information-theoretic approaches rely on the notion of entropy or uncertainty between sequences, whereas word-statistic approaches, which we focus on here, measure the probabilistic properties of 'words' within a single sequence. An advantage of word-statistics methods, which have been used previously for interspecies genetic analyses [14, 36, 15, 1, 8], is the generation of a one-to-one correspondence that does not rely on pairwise comparisons, so each sequence occupies a point in a genetic $\mathbb{R}^n$ space. One of the most widely used word-statistics are k-mer (also k-word) counts (kmc), which is the count of the occurrence of specific substrings (or words) formed by all possible combinations of length k of nucleotides in a genetic sequence; a more complete explanation can be found in the Materials and methods section. Alternatively, while k-mer counts ignore the positional information of words, a new method called natural vectors (NV) has been proposed that retains information on the general distribution of words by creating an array of values using summary statistics such as count, mean distance to the first site of the sequence, and the variance of that distance. This method has spawned several other ways to characterise genetic sequences [37, 6, 16, 5, 26], a more detailed explanation and exploration of which can be found in the Supplementary Material.

In general the preprocessing of the data to extract information that the classification and clustering algorithms can manipulate and analyse is known as feature extraction. We call the natural-vector based characterisation of sequences based on summary statistics, Natural Vector features (NVf). These features can be processed using dimensionality reduction and unsupervised classification algorithms, to help with the interpretability of relations among the sampled viruses by projecting the data onto lower dimensions. We assessed the potential for using alignment-free features, projected into lower dimensions using PacMAP, to identify clusters of sequences similar to the clustering obtained using phylogenetic methods, and particularly to pick out the VoCs. If successful, this may be a useful tool to help identify newly emerging and rapidly growing lineages quickly and without the need to generate massive phylogenies.

NVfs can also be processed in supervised classification algorithms to classify or assign sequences to already identified viral lineages. Pangolin [24], for example, uses a random forest algorithm to assign aligned sequences to previously designated lineages based on the patterns of SNPs across the genome. A potential problem with word-statistic methods is that they risk losing information that might be key to maintaining important distinctions between strongly similar samples, as could be the case for different lineages of the same viral population. We therefore explored the utility of different word-statistic features, and their application on different supervised Machine Learning algorithms, for describing global SARS-CoV-2 genetic diversity and the classification of known variants.

The approaches that we describe here have the potential to effectively designate and assign new sequences in a computationally efficient way, whilst enabling the easy incorporation new data. Although unlikely to replace current state-of-the-art phylogenetic approaches, they may prove to be important complementary tools to track the evolution of SARS-CoV-2 as well as other rapidly evolving pathogens.

Methods

The data was obtained from the GISAID database [30], approximately 250 000 high coverage SARS-CoV-2 sequences were downloaded (roughly 4% of $\sim$ 6 000 000 available at the time) from 2 July 2021 to 15 December 2021, providing a highly informative representation of VoC and other genomic information ascertained through time. The download was filtered to include only complete ($>$ 29 000 nucleotide) and high-coverage ($<$ 1% loci identified as N) sequences. Later on, each sequence was aligned to the official sequence hCov-19/Wuhan/WIV04/2019 [23] using the tool MAFFT [11] on an Ubuntu Windows subsystem for Linux (version 20.04.1 LTS) and Biopython version 1.78 scripts. Once aligned, a subset of nucleotides was extracted from the informative sites suggested by Lythgoe et al. [17]. This subset was then pre-processed as one-hot code, to later serve as input into selected supervised classification algorithms such as decision trees (dt), random forests (rf), k-nearest neighbors (knn) and support vector machines (svm) to produce an assignment of sequences into lineages as similar to the current "gold standard" as possible, where we use the implementations of methods provided by Scikit-learn [25]. The labels used as "ground truth" were obtained from the classification performed by Pangolin [24] version 1.2.101. Given that this tool provides two classifications labelled as "lineage" and "Scorpio call", and we decided to use the latter in order to focus on VoC.

In order to calculate features as in the main text and Supplementary Material, we wrote bespoke Python code to extract the required word statistic features. The most important such feature in our analysis is the k-mer count (kmc), which is produced by sliding a window of length k through the whole sequence shifting one position at the time obtaining $n-k+1$ overlapping strings, where n is the length of the sequence. From there, the occurrences of words of length k formed by all combinations of nucleotides are counted, resulting in a dictionary with the frequencies of each k-mer. This means that for a sequence $AAAAACGT$, this algorithm would extract a dictionary of 3-mer: $\{AAA:3, AAC:1, ACG:1, CGT:1\}$. For simplicity, the value k was set to 3 for all k-dependant algorithms. After the feature extraction was performed, an exploration of selected supervised classification algorithms was performed to evaluate performance-accuracy trade-offs with results as shown in Figure 4. The performance of the trained models was then evaluated using testing data to measure its accuracy (defined as: $(true\,positives + true\,negatives)/(total\,population)$).

The aforementioned methods of word-statistics to characterise the sequences were then projected onto low dimensions using graph-based dimensional reduction algorithms, as a first step towards unsupervised classification. Methods such as t-SNE [18] and UMAP [20] were explored, however we found that the more recent approach of PaCMAP [35] was the most robust for replication; results are obtained using version 0.5.2.1. This algorithm goes through three main steps: graph construction; initialisation of solution; and iterative optimisation. The graph construction is based on distinguishing three types of edges, neighbor pairs (NB): nearest neighbors from each observation, mid-near pairs (MN) and farther pairs (FP). Principal Component Analysis is used to initialise the solution. Finally, the iterative solution tries to minimise the loss function L_{PaCMAP}, defined as

$$L_{\text{PaCMAP}} = W_{\text{NB}} L_{\text{NB}} + W_{\text{MN}} L_{\text{MN}} + W_{\text{FP}} L_{\text{FP}},$$

where

$$L_{\text{NB}} = \sum_{i,j} \frac{\tilde{d}_{ij}}{10 + \tilde{d}_{ij}},$$

$$L_{\text{MN}} = \sum_{i,k} \frac{\tilde{d}_{ik}}{10\,000 + \tilde{d}_{ik}},$$

$$L_{\text{FP}} = \sum_{i,l} \frac{\tilde{d}_{il}}{1 + \tilde{d}_{il}},$$

$$\text{where } \tilde{d}_{ab} = \|y_a - y_b\|^2 + 1.$$

The pairs are formed between i, which is the base data-point; j, the nearest neighbor data-points; k.

which is the mid-near data-points; and l, the far data-points. Furthermore y is the array of coordinates of data-points in the original R^n graph, with the sub-index indicating a given subset.

To compare these novel analytics to the current state-of-the-art, a sub-sample limited to 1% of the total data-set, stratified by label, was processed to generate phylogenetic trees from the aligned sequences using the tool IQTree [22]. The volume of the sub-sample was limited to 1% due to the computational cost of processing high volumes of SARS-CoV-2 genetic sequences. The same sub-sample was projected using PaCMAP to visualise the contrast between the two techniques. This method has the advantage of being able to cope with orders of magnitude more sequences than the current options and offers more flexibility in the addition of new data.

To explore the manifolds emerging from the processing of features and dimensionality reduction, clustering algorithms such as HDBSCAN [19] (version 0.8.27) and the novel one CLASSIX [4] (version 0.6.5), were applied. HDBSCAN is an extension of DBSCAN, and finds clusters by assuming an existing density function from which the data-points are drawn, thus producing a cluster-tree assuming that each peak in said density function is a branch and then by "runt pruning" – focusing on branches with higher density – which is defined as groups of "core points", namely data-points having at least a predefined number of data-points within a neighborhood of a predefined radius and constraining them to not overlap. This algorithm proved to be fast and yielded far better results than standard methods such as k-means.

Additionally, we also tried the more recently proposed method CLASSIX, which has been optimised for low computational complexity and it is based on sorting the data-points based on the principal component, starting with the first point and assigning all the points within a predefined radius to the same group, then making the same with the next unassigned data-point. When all data-points have been assigned to a group, they are merged from clusters if their initial data-point of the group is found within a radius 1.5 times the aforementioned predefined distance. The results of this analysis including the whole data-set can be seen in Figure 2.

The scripts and a subsample of the dataset are available for replication at the repository:
https://github.com/robcah/LineageIdentificationByML/

Results

To demonstrate the potential of using alignment-free approaches to represent the genetic structure of SARS-Cov-2, we first compared a projection of kmc from unaligned sequences with a phylogeny generated using IQTree [22] using the same sequences, although previously aligned, see Figure 1. Because of the computational costs of generating phylogenetic trees, we used a sub-sample of 1% of $\sim$ 250 000 high coverage genomes downloaded from GISAID [30], namely 1% of undefined bases (N), and a minimal longitud of 29 000 nucleotides, between the dates of 2 July and 15 December 2021. The multi-dimensional kmc features were projected onto a 2-dimensional space using the non-linear dimension reduction tool PaCMAP [35] (detailed in Materials and methods), with each point representing the location of the sequence in the estimated genetic landscape. The points are coloured based on the Scorpio labelling by the Pangolin tool (see Materials and methods) [24]. The clustering of the significant VoCs such as Alpha (B.1.1.7), Delta (B.1.617.2), Omicron (BA.1), and Gamma (P.1), demonstrates the potential of the combination of word-statistics features and dimension reduction tools to produce interpretable structures for the distribution of variants in a genetic space.

To quantitatively measure the clusters formed by a 3-dimensional PaCMAP projection of the same data, we implemented two clustering algorithms, HDBSCAN [19] and CLASSIX [4] (Figure 2). The Scorpio labelling, produced by Pangolin, was used as a "ground truth" to calculate the adjusted Rand index (ARI) [28] and adjusted mutual information (AMI) [21] of the clustering matrices. The newer CLASSIX outperformed the more established HDBSCAN, with ARI and AMI values of 0.3198 and 0.5393 for CLASSIX, versus 0.1369 and 0.4431 for HDBSCAN. For both metrics, a value of 1 would mean the clusters are equal to the ground truth.

We were next interested in how the PacMAP projection of kmc counts performed when restricted to the major structural protein regions, Figure 3. To accurately identify the different gene regions we used the aligned sequences. For E (the gene encoding the envelope protein), we observed a large cluster with little or no differentiation among lineages, plus three clusters of minimal size, one containing mainly BA.1 (Omicron) sequences, and one containing mainly B.1.351 (Beta) sequences. Slightly more differentiation was seen for M (encoding the membrane), with Omicron (BA.1) and Delta (B.1.617.2) forming distinct clusters. These patterns become complex for N (encoding the nucleocapsid) and S (encoding the spike), although there is still some tendency for similar sequences to cluster together. The relative homogeneity observed for E and M may be due their relatively short lengths (227 and 668 nucleotides), and their relatively steady rates of evolution. We speculate the greater complexity seen for E and S may be a consequence of their longer lengths (1 259 and 3 821) and the unique "constellations" of mutations that define the different VoCs, combined with convergent evolution.

Finally, we compared the performance of different word-statistic features and ML algorithms to classify sequences using Scorpio classifications as the ground truth, and with accuracy and processing time the key indicators Figure 4. Machine Learning (ML) algorithms are computational methods that build models using training data to make predictions, such as classification, without those models needing to be explicitly programmed.

The features we considered were kmc, natural vectors (NV), and informative sites (loci; these are known lineage defining mutations). The ML algorithms that we considered were: (i) decision trees (dt), which make a classification after a series of decision nodes based on thresholds for the different attributes of the datapoints; (ii) random forests (rf), i.e. ensembles of multiple smaller decision trees which learn to classify by training with a subsample of the attributes and datapoints, which then "vote" on the classification outcome; (iii) k-nearest neighbors (knn), i.e. classification obtained by measuring the distance to a number k of nearest neighboring datapoints; (iv) support vector machines (svm), which delimit "borders" within the space in which the datapoints exist to then produce the classification according to the location of the datapoint to classify.

The use of decision trees and random forest using informative sites (loci-dt, loci-rf) demonstrated a good balance of accuracy and processing cost in comparison to other relatively successful features-algorithm pairs like KMNV-dt and kmc-rf. The algorithm with the highest accuracy was svm but was computationally expensive. The pair kmc-knn showed smaller costs and reasonable accuracy. Although using informative sites came out as reliable and inexpensive, it is important to bear in mind there is a significant preprocessing in the curation of informative sites and the alignment of sequences, which is not required for the other feature extraction methods.

Discussion

The unprecedented amount of genetic data generated during the pandemic demands the development of methods to analyse it thoroughly, with fluidity and efficiency. Without showing a benefit to curating this data in the future there is a risk that it will be removed or deleted. The application of algorithms that characterise full sequences and project them onto low dimensional space has the potential to reveal genetic relationships among a huge number of sequences. The supervision of these genetic spaces with automated clustering algorithms could provide alternative methods to discover niches of specific viral variants. Additionally, this may enable the early discovery of, or add supporting evidence to, the emergence of new VoCs. Whilst phylogenetics remains the gold-standard for understanding the genetic topology of viral populations, the proposed methods have the advantage of being able to manage several orders of magnitude more sequences than the current methods.

Future research along this line could deepen on the optimal values of k for a proper extraction of information from the sequences. Additionally, there could be an evaluation on which features and combinations of them might be more significant for a more reliable construction of genetic spaces. In alternative, the application of other alignment-free methods such as protein NV [39, 32], graphical representation [32], and Fourier power spectrum [9], among others, could add value to these analy-

sis. Furthermore, previous methods for phylogenetic reconstruction using non-Euclidean metrics, like Kullback-Leibler, Yaus-Hausdorff, or Structural Similarity Index Measure, applying them to dimension reduction algorithms might generate more accurate representations of the genetic landscape. The presented results show values of ARI and AMI that might seem low, albeit encouraging; however, these metrics can be considered optimal for two cases [29]: ARI for when clustering present equally sized clusters, and AMI for when the clustering is highly unbalanced, thus making the AMI index more meaningful for the used dataset. Further analysis, like combination of features, could improve these metrics and reduce noise.

The methods presented in this paper provide the means to accelerate the identification of new pathogen variants of concern by applying alignment-free methods based on word-statistics. However, this research avenue is still in its very early stages. Further exploration of hyperparameters, mixing of different methods of characterisation, and data preprocessing, could help to enhance their accuracy and reliability.

Acknowledgements

This project was possible thanks to the support of the Joint Universities Pandemic and Epidemiological Research Consortium (JUNIPER – https://maths.org/juniper/), the Engineering and Physical Sciences Research Council, the Wellcome Trust and the Li Ka Shing Foundation. LP gratefully acknowledges the Wellcome Trust and Royal Society (grant number 202562/Z/16/Z). TAH is supported by the Royal Society (grant number INF\R2\180067). IH is supported by the National Institute for Health Research Health Protection Research Unit (NIHR HPRU) in Emergency Preparedness and Response and the National Institute for Health Research Policy Research Programme in Operational Research (OPERA). The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR, the Department of Health or Public Health England. LP, TH and IH are also supported by the JUNIPER modelling consortium (grant number MR/V038613/1) and by the Alan Turing Institute for Data Science and Artificial Intelligence under the EPSRC grants EP/N510129/1 and EP/V027468/1.

Figures

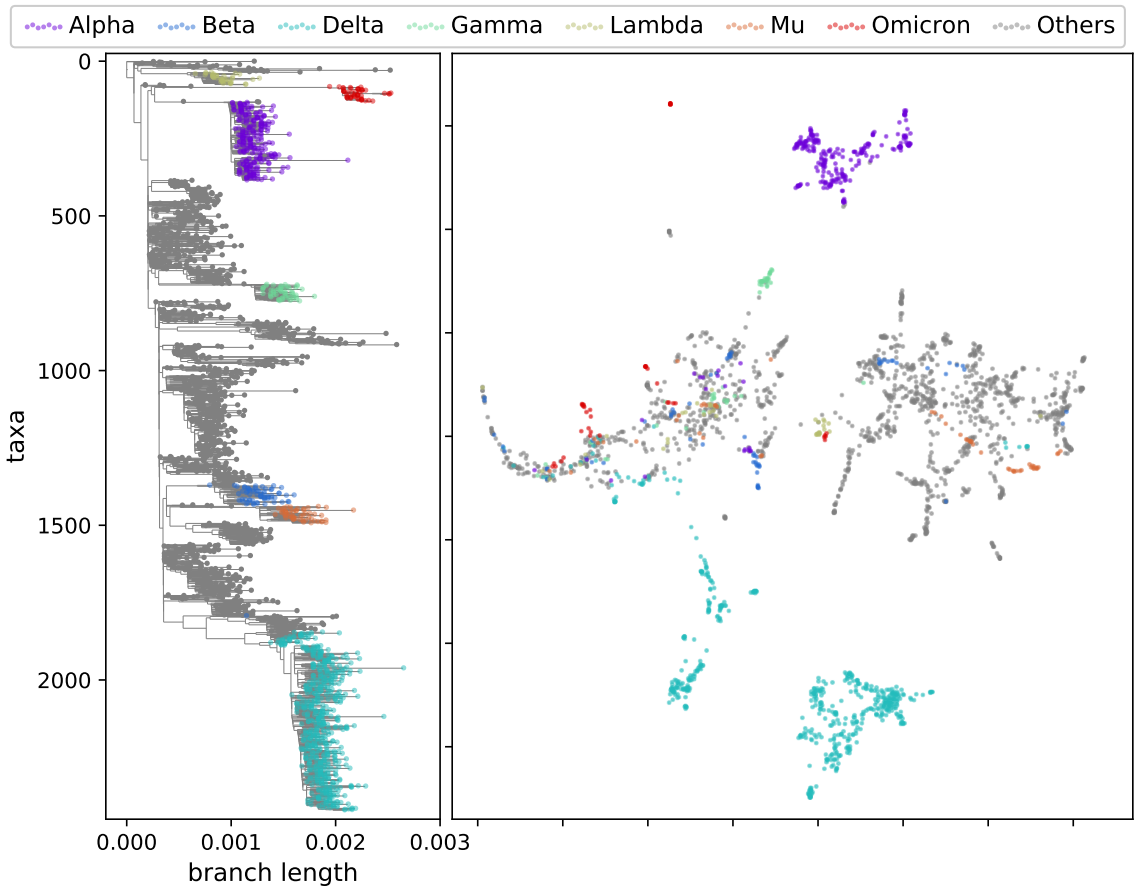


Figure 1: Comparison of a phylogenetic tree generated using IQTree (left) and PaCMAP projection of k-mer counts (right). This plot was generated using a sub-sample of 1% of the high coverage genomes (1% of undefined bases with a longitud of $\sim 29K$ nucleotides) available on GISAID [30], downloaded between the dates of 02/07/2021 to 15/12/2021, due to high processing cost of phylogenetic-tree generation. This shows the potential of the combination of word-statistics features and dimension reduction tools to structure the distribution of variants in a genetic space.

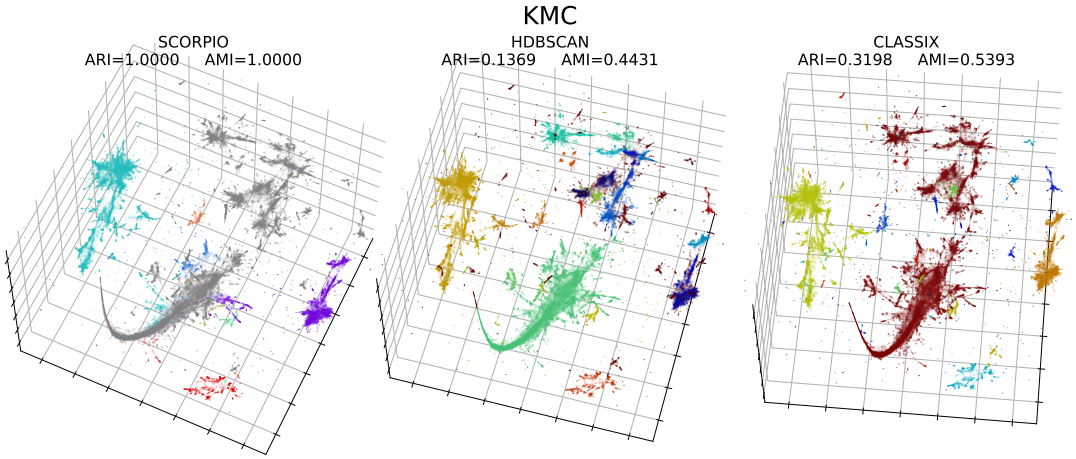


Figure 2: 3d PaCMAP projection and clustering analysis of genetic topology generated using kmc (left) colorised by Scorpio labelling, same color code as Figure 1, (center) clustering by HDBSCAN, (right) clustering by the novel CLASSIX algorithm. The formation of well defined clusters show the potential of these tools to help us gain insight on phylogenetical relations of vast amounts of sequences in an relatively resource inexpensive fashion.

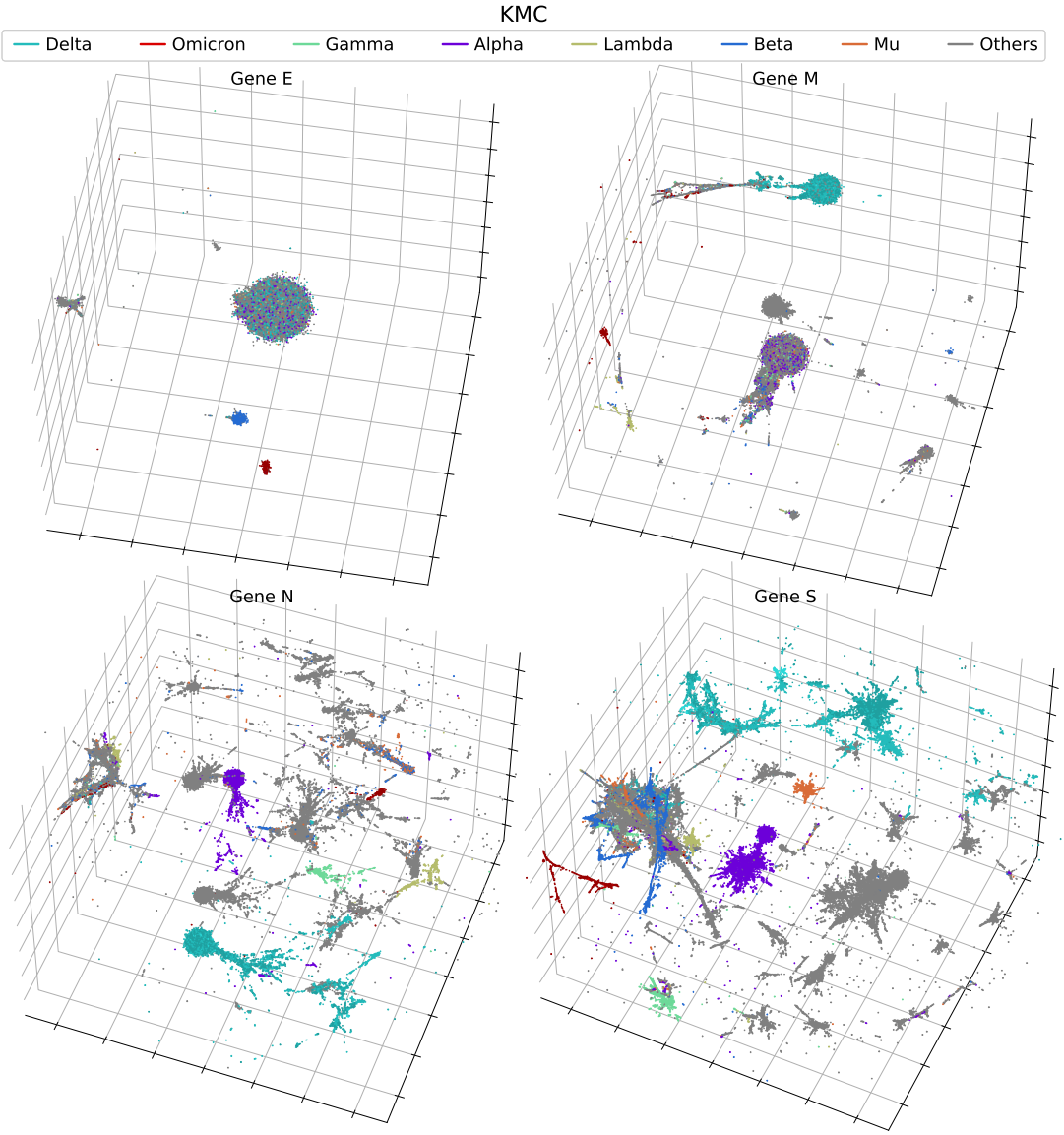


Figure 3: 3d PaCMAP projection of specific structural protein regions, generated using kmc and coloured by Scorpio labelling. The formation of clusters allows us to see the contribution of the different proteins to the differentiation within the viral population.

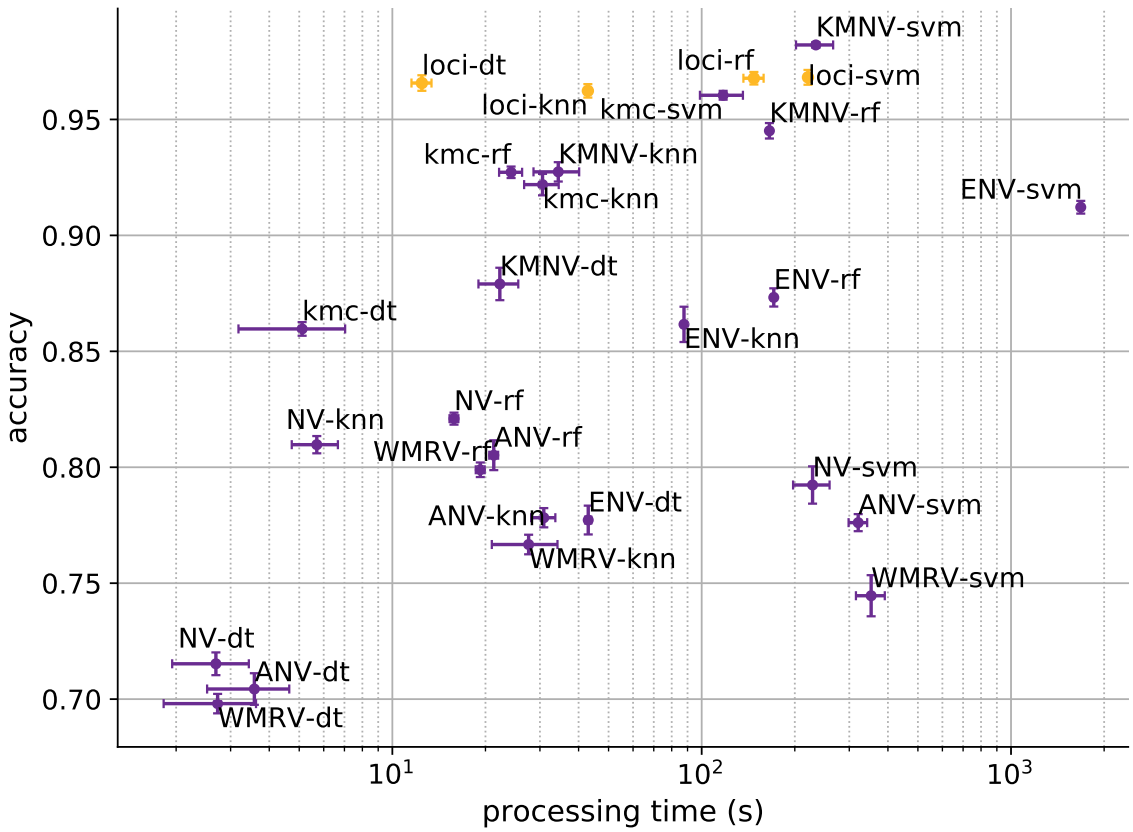


Figure 4: Summary of trade-off by pairs feature-algorithm. Features consist of: informative sites (loci), natural vector (NV), accumulated NV (ANV), degenerate bases NV (WMRV), k-mer NV (NV), and extended NV (ENV). Moreover, the machine learning algorithms explored were: decision tree (dt), random forest (rf), k-nearest neighbors (knn), and support vector machines (svm). The error bars represent the bootstrapping in which 95% of the values were found. Yellow values are taken as a state-of-the-art benchmark. For these results, a sub-sample of 20% of the whole data ($\sim$ 250 000 sequences) was used to train, while 5% for testing accuracy.

References

[1] E. S. Allman, J. A. Rhodes, and S. Sullivant. Statistically Consistent k-mer Methods for Phylogenetic Tree Reconstruction. *Journal of Computational Biology*, 24(2):153–171, 2017.

[2] D. Alvarez-Ponce, P. Lopez, E. Bapteste, and J. O. McInerney. Gene similarity networks provide tools for understanding eukaryote origins and evolution. *Proceedings of the National Academy of Sciences*, 110(17):E1594–E1603, 2013. _eprint: https://www.pnas.org/doi/pdf/10.1073/pnas.1211371110.

[3] Y. Cao, A. Yisimayi, F. Jian, W. Song, T. Xiao, L. Wang, S. Du, J. Wang, Q. Li, X. Chen, Y. Yu, P. Wang, Z. Zhang, P. Liu, R. An, X. Hao, Y. Wang, J. Wang, R. Feng, H. Sun, L. Zhao, W. Zhang, D. Zhao, J. Zheng, L. Yu, C. Li, N. Zhang, R. Wang, X. Niu, S. Yang, X. Song, Y. Chai, Y. Hu, Y. Shi, L. Zheng, Z. Li, Q. Gu, F. Shao, W. Huang, R. Jin, Z. Shen, Y. Wang, X. Wang, J. Xiao, and X. S. Xie. BA.2.12.1, BA.4 and BA.5 escape antibodies elicited by Omicron infection. *Nature*, June 2022.

[4] X. Chen and S. Güttel. Fast and explainable clustering based on sorting. 2022. Publisher: arXiv Version Number: 1.

[5] M. Deng, C. Yu, Q. Liang, R. L. He, and S. S.-T. Yau. A novel method of characterizing genetic sequences: Genome space with biological distance and applications. *PLoS ONE*, 6(3):e17293, Mar. 2011. Publisher: Public Library of Science (PLoS).

[6] R. Dong, L. He, R. L. He, and S. S.-T. Yau. A novel approach to clustering genome sequences using inter-nucleotide covariance. *Frontiers in Genetics*, 10, Apr. 2019. Publisher: Frontiers Media SA.

[7] J. Hadfield, C. Megill, S. M. Bell, J. Huddleston, B. Potter, C. Callender, P. Sagulenko, T. Bedford, and R. A. Neher. Nextstrain: real-time tracking of pathogen evolution. *Bioinformatics*, 34(23):4121–4123, Dec. 2018.

[8] K. Hatje and M. Kollmar. A Phylogenetic Analysis of the Brassicales Clade Based on an Alignment-Free Sequence Comparison Method. *Frontiers in Plant Science*, 3, 2012. Publisher: Frontiers Media SA.

[9] T. Hoang, C. Yin, H. Zheng, C. Yu, R. Lucy He, and S. S.-T. Yau. A new method to cluster DNA sequences using Fourier power spectrum. *Journal of Theoretical Biology*, 372:135–145, May 2015.

[10] M. Kaden, K. S. Bohnsack, M. Weber, M. Kud\la, K. Gutowska, J. Blazewicz, and T. Villmann. Learning vector quantization as an interpretable classifier for the detection of SARS-CoV-2 types based on their RNA sequences. *Neural Computing and Applications*, Apr. 2021. Publisher: Springer Science and Business Media LLC.

[11] K. Katoh. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Research*, 30(14):3059–3066, July 2002. Publisher: Oxford University Press (OUP).

[12] M. Kokot, M. Długosz, and S. Deorowicz. KMC 3: counting and manipulating k-mer statistics. *Bioinformatics*, 33(17):2759–2761, May 2017. _eprint: https://academic.oup.com/bioinformatics/article-pdf/33/17/2759/25163903/btx304.pdf.

[13] D. Lebatteux, A. M. Remita, and A. B. e. Diallo. Toward an Alignment-Free Method for Feature Extraction and Accurate Classification of Viral Sequences. *Journal of Computational Biology*, 26(6):519–535, June 2019. Publisher: Mary Ann Liebert Inc.

[14] C. Li, Y. Yang, M. Jia, Y. Zhang, X. Yu, and C. Wang. Phylogenetic analysis of DNA sequences based on -word and rough set theory. *Physica A: Statistical Mechanics and its Applications*, 398:162–171, Mar. 2014. Publisher: Elsevier BV.

[15] Y. Li, L. He, R. L. He, and S. S.-T. Yau. Zika and Flaviviruses Phylogeny Based on the Alignment-Free Natural Vector Method. *DNA and Cell Biology*, 36(2):109–116, Feb. 2017. Publisher: Mary Ann Liebert Inc.

[16] Y. Li, L. He, R. Lucy He, and S. S.-T. Yau. A novel fast vector method for genetic sequence comparison. *Scientific Reports*, 7(1), Sept. 2017. Publisher: Springer Science and Business Media LLC.

[17] K. Lythgoe and e. al. Informative sites. *Unpublished*, 2021. Published: personal communication.

[18] L. v. d. Maaten and G. Hinton. Visualizing Data using t-SNE. *Journal of Machine Learning Research*, 9(86):2579–2605, 2008.

[19] L. McInnes and J. Healy. Accelerated hierarchical density based clustering. In *2017 IEEE International Conference on Data Mining Workshops (ICDMW)*. IEEE, Nov. 2017.

[20] L. McInnes, J. Healy, and J. Melville. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction, 2018.

[21] M. Meilă. Comparing clusterings—an information based distance. *Journal of Multivariate Analysis*, 98(5):873–895, 2007.

[22] L.-T. Nguyen, H. A. Schmidt, A. von Haeseler, and B. Q. Minh. IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. *Molecular Biology and Evolution*, 32(1):268–274, Nov. 2014. _eprint: https://academic.oup.com/mbe/article-pdf/32/1/268/13171186/msu300.pdf.

[23] P. Okada, R. Buathong, S. Phuygun, T. Thanadachakul, S. Parnmen, W. Wongboot, S. Waicharoen, S. Wacharapluesadee, S. Uttayamakul, A. Vachiraphan, M. Chittaganpitch, N. Mekha, N. Janejai, S. Iamsirithaworn, R. T. Lee, and S. Maurer-Stroh. Early transmission patterns of coronavirus disease 2019 (COVID-19) in travellers from Wuhan to Thailand, January 2020. *Eurosurveillance*, 25(8), Feb. 2020. Publisher: European Centre for Disease Control and Prevention (ECDC).

[24] \. A. O'Toole, E. Scher, A. Underwood, B. Jackson, V. Hill, J. T. McCrone, R. Colquhoun, C. Ruis, K. Abu-Dahab, B. Taylor, C. Yeats, L. Du Plessis, D. Maloney, N. Medd, S. W. Attwood, D. M. Aanensen, E. C. Holmes, O. G. Pybus, and A. Rambaut. Assignment of Epidemiological Lineages in an Emerging Pandemic Using the Pangolin Tool. *Virus Evolution*, July 2021. Publisher: Oxford University Press (OUP).

[25] 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 Duchesnay. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12(85):2825–2830, 2011.

[26] S. Pei, W. Dong, X. Chen, R. L. He, and S. S.-T. Yau. Fast and accurate genome comparison using genome images: The Extended Natural Vector Method. *Molecular Phylogenetics and Evolution*, 141:106633, Dec. 2019. Publisher: Elsevier BV.

[27] A. Rambaut, E. C. Holmes, \. A. O'Toole, V. Hill, J. T. McCrone, C. Ruis, L. du Plessis, and O. G. Pybus. A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. *Nature Microbiology*, 5(11):1403–1407, July 2020. Publisher: Springer Science and Business Media LLC.

[28] W. M. Rand. Objective Criteria for the Evaluation of Clustering Methods. *Journal of the American Statistical Association*, 66(336):846–850, 1971. Publisher: Taylor & Francis _eprint: https://www.tandfonline.com/doi/pdf/10.1080/01621459.1971.10482356.

[29] S. Romano, N. X. Vinh, J. Bailey, and K. Verspoor. Adjusting for Chance Clustering Comparison Measures. *J. Mach. Learn. Res.*, 17(1):4635–4666, Jan. 2016. Publisher: JMLR.org.

[30] Y. Shu and J. McCauley. GISAID: Global initiative on sharing all influenza data – from vision to reality. *Eurosurveillance*, 22(13), Mar. 2017. Publisher: European Centre for Disease Control and Prevention (ECDC).

[31] H. Tegally, M. Moir, J. Everatt, M. Giovanetti, C. Scheepers, E. Wilkinson, K. Subramoney, S. Moyo, D. G. Amoako, C. Baxter, C. L. Althaus, U. J. Anyaneji, D. Kekana, R. Viana, J. Giandhari, R. J. Lessells, T. Maponga, D. Maruapula, W. Choga, M. Matshaba, S. Mayaphi, N. Mbhele, M. B. Mbulawa, N. Msomi, Y. Naidoo, S. Pillay, T. J. Sanko, J. E. San, L. Scott, L. Singh, N. A. Magini, P. Smith-Lawrence, W. Stevens, G. Dor, D. Tshiabuila, N. Wolter, W. Preiser, F. K. Treurnicht, M. Venter, M. Davids, G. Chiloane, A. Mendes, C. McIntyre, A. O'Toole, C. Ruis, T. P. Peacock, C. Roemer, C. Williamson, O. G. Pybus, J. Bhiman, A. Glass, D. P. Martin, A. Rambaut, S. Gaseitsiwe, A. v. Gottberg, and T. de Oliveira. Continued Emergence and Evolution of Omicron in South Africa: New BA.4 and BA.5 lineages. *medRxiv*, 2022.

[32] K. Tian, X. Yang, Q. Kong, C. Yin, R. L. He, and S. S.-T. Yau. Two Dimensional Yau-Hausdorff Distance with Applications on Comparison of DNA and Protein Sequences. *PLOS ONE*, 10(9):e0136577, Sept. 2015.

[33] A. Tuekprakhon, R. Nutalai, A. Dijokaite-Guraliuc, D. Zhou, H. M. Ginn, M. Selvaraj, C. Liu, A. J. Mentzer, P. Supasa, H. M. E. Duyvesteyn, R. Das, D. Skelly, T. G. Ritter, A. Amini, S. Bibi, S. Adele, S. A. Johnson, B. Constantinides, H. Webster, N. Temperton, P. Klenerman, E. Barnes, S. J. Dunachie, D. Crook, A. J. Pollard, T. Lambe, P. Goulder, N. G. Paterson, M. A. Williams, D. R. Hall, C. Conlon, A. Deeks, J. Frater, L. Frending, S. Gardiner, A. Jämsén, K. Jeffery, T. Malone, E. Phillips, L. Rothwell, L. Stafford, E. E. Fry, J. Huo, J. Mongkolsapaya, J. Ren, D. I. Stuart, and G. R. Screaton. Antibody escape of SARS-CoV-2 Omicron BA.4 and BA.5 from vaccine and BA.1 serum. *Cell*, 185(14):2422–2433.e13, 2022.

[34] S. Vinga and J. Almeida. Alignment-free sequence comparison–a review. *Bioinformatics*, 19(4):513–523, Mar. 2003. Publisher: Oxford University Press (OUP).

[35] Y. Wang, H. Huang, Cynthia Rudin, and Y. Shaposhnik. Understanding How Dimension Reduction Tools Work: An Empirical Approach to Deciphering t-SNE, UMAP, TriMap, and PaCMAP for Data Visualization. *Journal of Machine Learning Research*, 22(201):1–73, 2021.

[36] Y. Wang, K. Tian, and S. S.-T. Yau. Protein Sequence Classification Using Natural Vector and Convex Hull Method. *Journal of Computational Biology*, 26(4):315–321, Apr. 2019. Publisher: Mary Ann Liebert Inc.

[37] J. Wen, R. H. Chan, S.-C. Yau, R. L. He, and S. S. Yau. K-mer natural vector and its application to the phylogenetic analysis of genetic sequences. *Gene*, 546(1):25–34, Aug. 2014. Publisher: Elsevier BV.

[38] C. Yu, S.-Y. Cheng, R. L. He, and S. S.-T. Yau. Protein map: An alignment-free sequence comparison method based on various properties of amino acids. *Gene*, 486(1-2):110–118, Oct. 2011. Publisher: Elsevier BV.

[39] C. Yu, M. Deng, S.-Y. Cheng, S.-C. Yau, R. L. He, and S. S.-T. Yau. Protein space: A natural method for realizing the nature of protein universe. *Journal of Theoretical Biology*, 318:197–204, Feb. 2013.

[40] A. Zielezinski, H. Z. Girgis, G. Bernard, C.-A. Leimeister, K. Tang, T. Dencker, A. K. Lau, S. R\" ohling, J. Choi, M. S. Waterman, M. Comin, S.-H. Kim, S. Vinga, J. S. Almeida, C. X. Chan, B. T. James, F. Sun, B. Morgenstern, and W. M. Karlowski. Benchmarking of alignment-free sequence comparison methods. Technical report, Cold Spring Harbor Laboratory, Apr. 2019.

[41] A. Zielezinski, S. Vinga, J. Almeida, and W. M. Karlowski. Alignment-free sequence comparison: benefits, applications, and tools. *Genome Biology*, 18(1), Oct. 2017. Publisher: Springer Science and Business Media LLC.

Supplementary Material

S.1 Supplementary Text

The viral evolution of SARS-CoV-2 in combination to the modern sequencing technologies presented a challenge for the quick analysis of the massive volume of genetic data. As mention in the main article, this project hypothesised that given that word-statistic methods were proven methods to create phylogenetical trees, the emergence of analogous structures could be appreciate through dimension reduction, producing an interpretable glance of the emerging manifold showing a distribution of variants in the genetic space of the likes of similarity networks [2].

The exploration of linear algorithms for dimensionality reduction, such as PCA, ICA or SVD did not yield results with sufficient interpretability. Therefore, other non-linear algorithms were used, such as t-SNE [18] and UMAP [20]. These two had a better result on the representation of the data, however the replicability was challenging since they seemed to be very sensible to the hyperparameter settings. The novel PaCMAP presented robust replicability, therefore was used to perform the analysis of this project.

Furthermore, from all the summary-statistics features explained in the main paper only KMC [12] results are shown, however we also implemented and tested more complex methods based on natural vectors (NV), which we detail and present below.

The NV approach was proposed by [5], who used standardised central moments to characterise the nucleotide distribution in each sequence producing a 12-dimensional vector. The NV of a genetic sequence $S = (s_1, s_2, ...s_n)$ of length n, where $s_i \in \{A, C, G, T\}, i \in \{1, \ldots, n\}$, is constituted by the following elements for each $\omega \in \{A, C, G, T\}$:

$$\begin{aligned} n_\omega &= \sum_{i=0}^{n} w_\omega(s_i); \\ \mu_\omega &= \sum_{i=0}^{n} i \frac{w_\omega(s_i)}{n_\omega}; \\ D_2^\omega &= \sum_{i=0}^{n} \frac{(i-\mu_\omega)^2 w_\omega(s_i)}{n_\omega n}. \end{aligned} \tag{S1}$$

Here the count of nucleotide ω, n_ω, the mean distance of said nucleotide to the origin of the sequence, μ_ω, and the variance of said distance, D_2^ω, are defined in terms of the indicator functions

$$w_\omega(s) = \begin{cases} 1 & \text{if } s = \omega, \\ 0 \text{ otherwise.} \end{cases}$$

Since we have a copy of the variables in Equation (S1) for each of the four amino acids, this leads to a 12-dimensional natural vector representation of the sequence S that can be written explicitly as

$$\mathrm{NV}(S) = (n_A, \mu_A, D_2^A, n_C, \mu_C, D_2^C, n_G, \mu_G, D_2^G, n_T, \mu_T, D_2^T)$$

Another characterisation of sequences comes from Accumulated Natural Vectors (ANV) [6] which forms its arrays by n_ω as defined in Equation (S1), together with average distance to the end of the sequence, ζ_ω and the covariance of the accumulation of nucleotides through S. These new elements can be defined as follows:

$$\begin{aligned} \zeta_\omega &= n_\omega(n + 1 - \mu_\omega); \\ \mathrm{cov}(\omega, \varpi) &= \sum_{i=1}^{n} i \frac{(\hat{u}_\omega(i) - \theta_\omega)(\hat{u}_\varpi(i) - \theta_\varpi)}{n_\omega n_\varpi}, \\ \text{where} \quad \theta_\omega &= \sum_{i=1}^{n} \frac{\hat{u}_\omega(i)}{n}. \end{aligned}$$

Here, $\hat{u}_\omega$ is the accumulation indication function, which is an array of the accumulation of nucleotide ω in position i, so for $S = AAGCGT$, $\hat{u}_A = (1, 2, 2, 2, 2, 2)$, for $\hat{u}_G = (0, 0, 1, 1, 2, 2)$ and so on for all nucleotides, or formally in terms of the indicator function

$$\hat{u}_\omega(i) = \sum_{j \leq i} w_\omega(s_j).$$

The covariance is calculated for all pairs of elements of $\{A, C, G, T\}$, which results in the 18-dimension array

$$\begin{aligned}\text{ANV}(S) =&(n_A, \zeta_A, n_C, \zeta_C, n_G, \zeta_G, n_T, \zeta_T, \text{cov}(A, A),\\ &\text{cov}(A, C), ..., \text{cov}(T, G), \text{cov}(T, T)).\end{aligned}$$

A different method called weak-amino-purine vector (or WMRV) [16] characterises the sequence S by producing three mappings of the sequence codifying it into three pairs of complementary bases. The first coding is onto purine (R) or pyrimidine (Y),

$$\text{cod}_{RY}(A) = \text{cod}_{RY}(G) = R, \quad \text{cod}_{RY}(C) = \text{cod}_{RY}(T) = Y.$$

The second coding is onto amino (M) or keto (K),

$$\text{cod}_{MK}(A) = \text{cod}_{MK}(C) = M, \quad \text{cod}_{MK}(G) = \text{cod}_{MK}(T) = K.$$

The third coding is onto strong H bond (S) or weak H bond (W),

$$\text{cod}_{SW}(C) = \text{cod}_{SW}(G) = S, \quad \text{cod}_{SW}(A) = \text{cod}_{SW}(T) = W.$$

And so for $S = AAGCGT$, $\text{cod}_{RY}(S) = RRRYRY$, $\text{cod}_{MK}(S) = MMKMKK$ and $\text{cod}_{SW}(S) = WWSSSW$. The resulting array has the following 18-dimensions:

$$\begin{aligned}\text{WMRV}(S) =&(n_R, \mu_R, D_2^R, n_Y, \mu_Y, D_2^Y, n_M, \mu_M,\\ &D_2^M, n_K, \mu_K, D_2^K, n_W, \mu_W, D_2^W, n_S, \mu_S, D_2^S).\end{aligned}$$

Another idea from [37] attempted to combine the use of k-mers and NV by characterising k-mer distributions within S. For this a window of length k is slid throughout the whole sequence, and the count, mean distance, and variance of the k-mers are characterised with the summary statistics in Equation (S1) for NV applied to each k-mer instead of each nucleotide. i.e. with $\omega \in \{AAA, AAC, ..., TTG, TTT\}$. This is called k-mer natural vectors, or KMNV, for $k = 3$.

Finally, the extended natural vector (ENV) [26] combines the frequency chaos game representation (FCGR) characterisation with the NV approach. In this method the objects to characterise are the discrete values $\omega \in \{0, ...255\}$ which represent the intensity of a pixel in the 2-dimensional FCGR, expressed as the square matrix Q with values $q(i, j) \in \omega$. The equations from NV are adapted to 2-dimensions as below:

$$\begin{aligned}\mu_{1,\omega} &= \frac{\sum_{s=1}^{n_\omega} i_{s,\omega}}{n_k},\\ \mu_{2,\omega} &= \frac{\sum_{s=1}^{n_\omega} j_{s,\omega}}{n_k},\\ D_{\omega,r,s} &= \sum_{t=1}^{n_k} \frac{(i_{t,\omega} - \mu_{1,\omega})^r \cdot (j_{t,k} - \mu_{2,k})^s}{n_k^{r+s} \cdot N^{r+s-1}},\end{aligned}$$

for pairs $(r, s) \in \{(1, 1), (0, 2), (2, 0)\}$. Therefore the resulting 1536-dimensional array is formed by the following values:

$$\begin{aligned}\text{ENV}(S) =&(n_0, \mu_{1,0}, \mu_{2,0}, D_{0,2,0}, D_{0,0,2}, D_{0,1,1}, ...,\\ &n_{255}, \mu_{1,255}, , \mu_{2,255}, D_{255,2,0}, D_{255,0,2}, D_{255,1,1})\end{aligned}$$

As seen in the following features there is mainly tendency for these characterisations to produce reasonable clusters for similar sequences. The overall projections of these natural-vectors features can be seen in Figures S3, S1, S2, S4, S5. While of interest, these do not provide qualitatively different insights from the more interpretable kmc feature extraction used in the main text.

For further completeness, the scripts and a subsample of the dataset are available for replication at the repository: https://github.com/robcah/LineageIdentificationByML/

S.2 Supplementary Figures

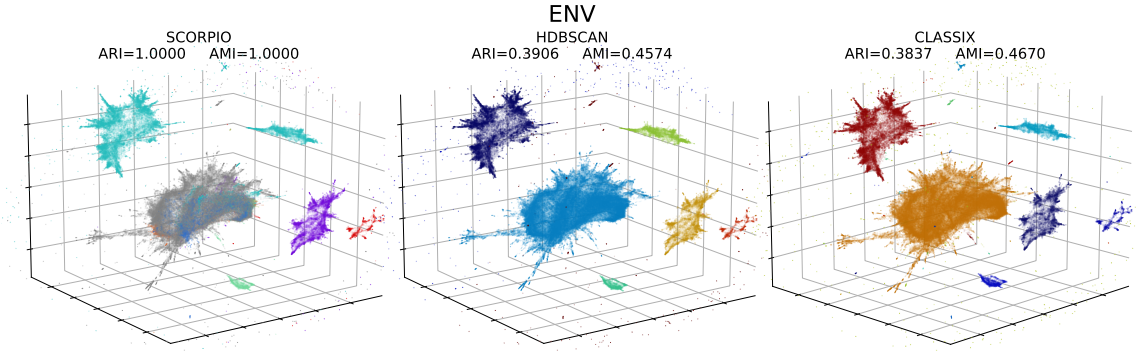


Figure S1: PaCMAP projection of the dataset extracting the ENV features [26], (left) colouring based on Scorpio labelling [24], (centre) unsupervised clustering identification using HDBSCAN [19], (right) unsupervised clustering identification using CLASSIX [4]

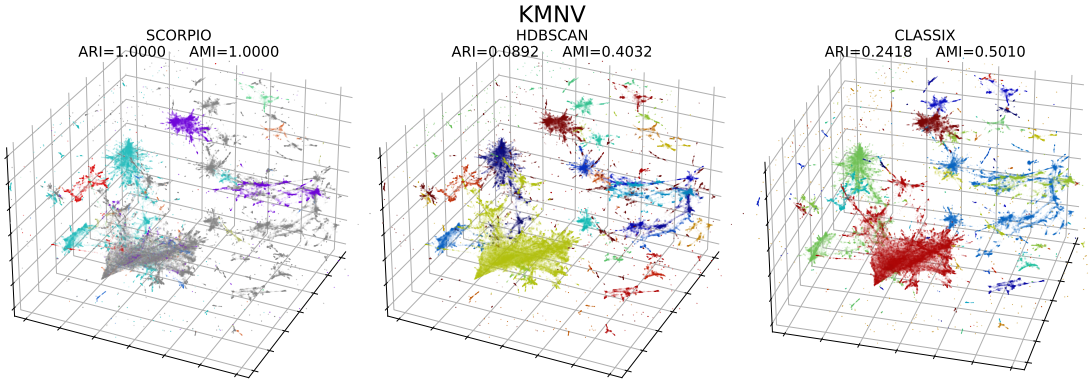


Figure S2: PaCMAP projection of the dataset extracting the KMNV features [37], (left) colouring based on Scorpio labelling [24], (centre) unsupervised clustering identification using HDBSCAN [19], (right) unsupervised clustering identification using CLASSIX [4]

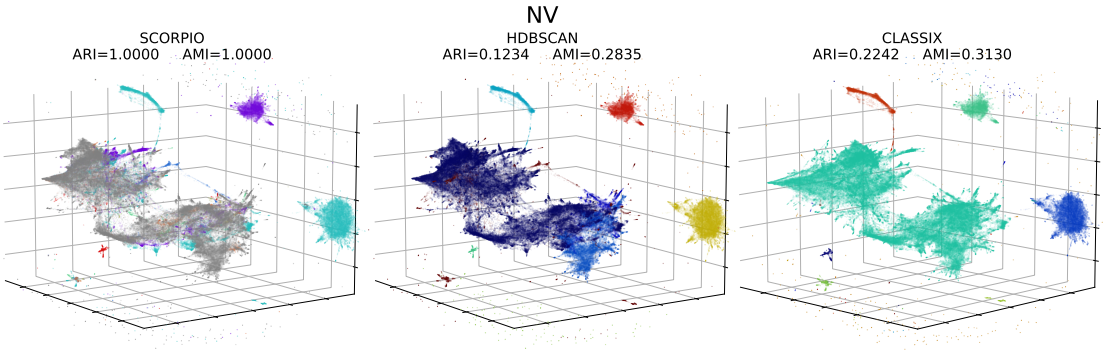


Figure S3: PaCMAP projection of the dataset extracting the NV features [5], (left) colouring based on Scorpio labelling [24], (centre) unsupervised clustering identification using HDBSCAN [19], (right) unsupervised clustering identification using CLASSIX [4]

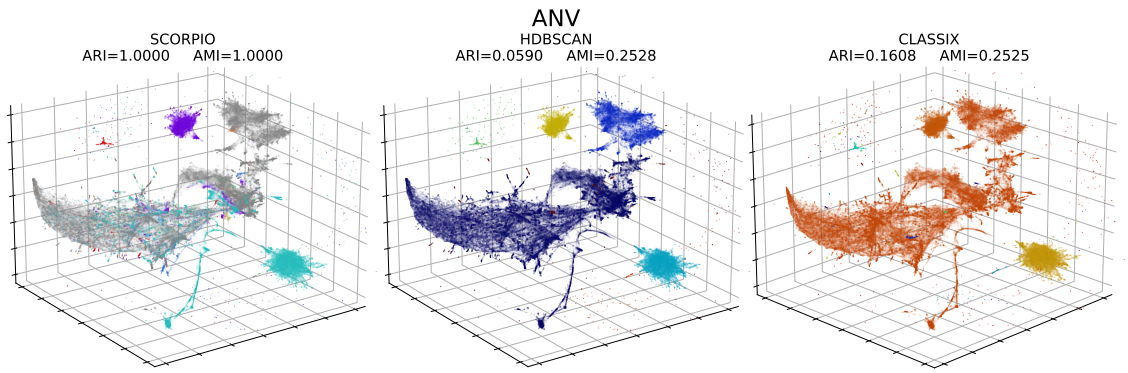


Figure S4: PaCMAP projection of the dataset extracting the ANV features [6], (left) colouring based on Scorpio labelling [24], (centre) unsupervised clustering identification using HDBSCAN [19], (right) unsupervised clustering identification using CLASSIX [4]

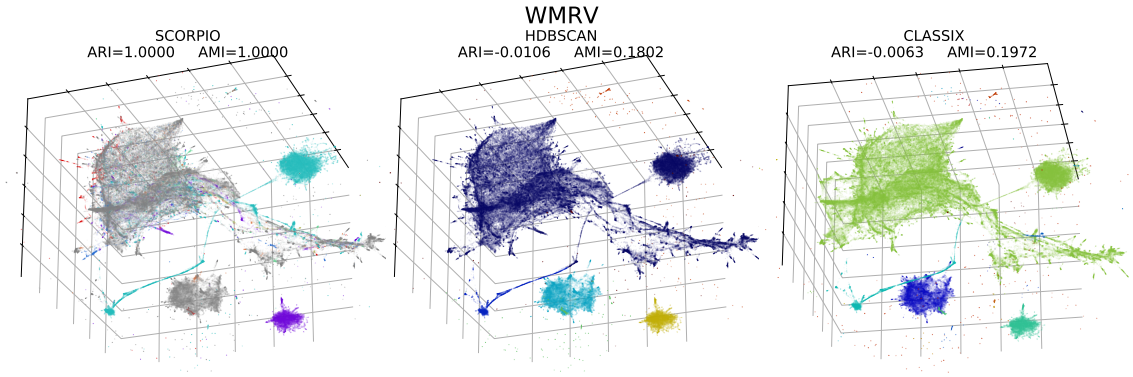
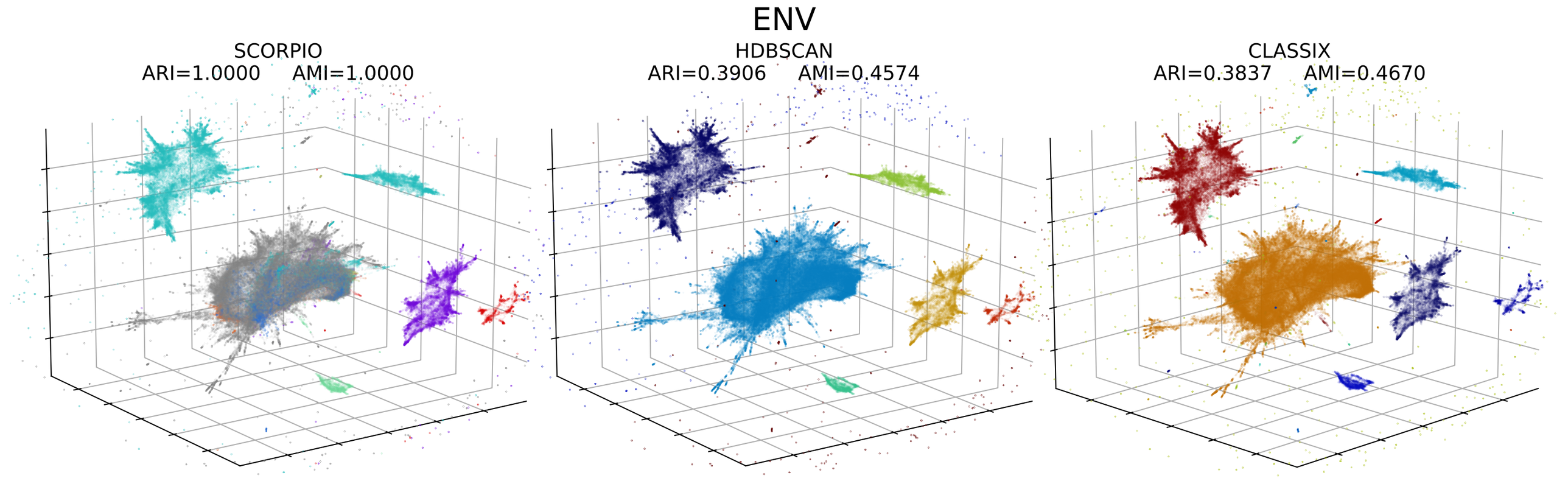


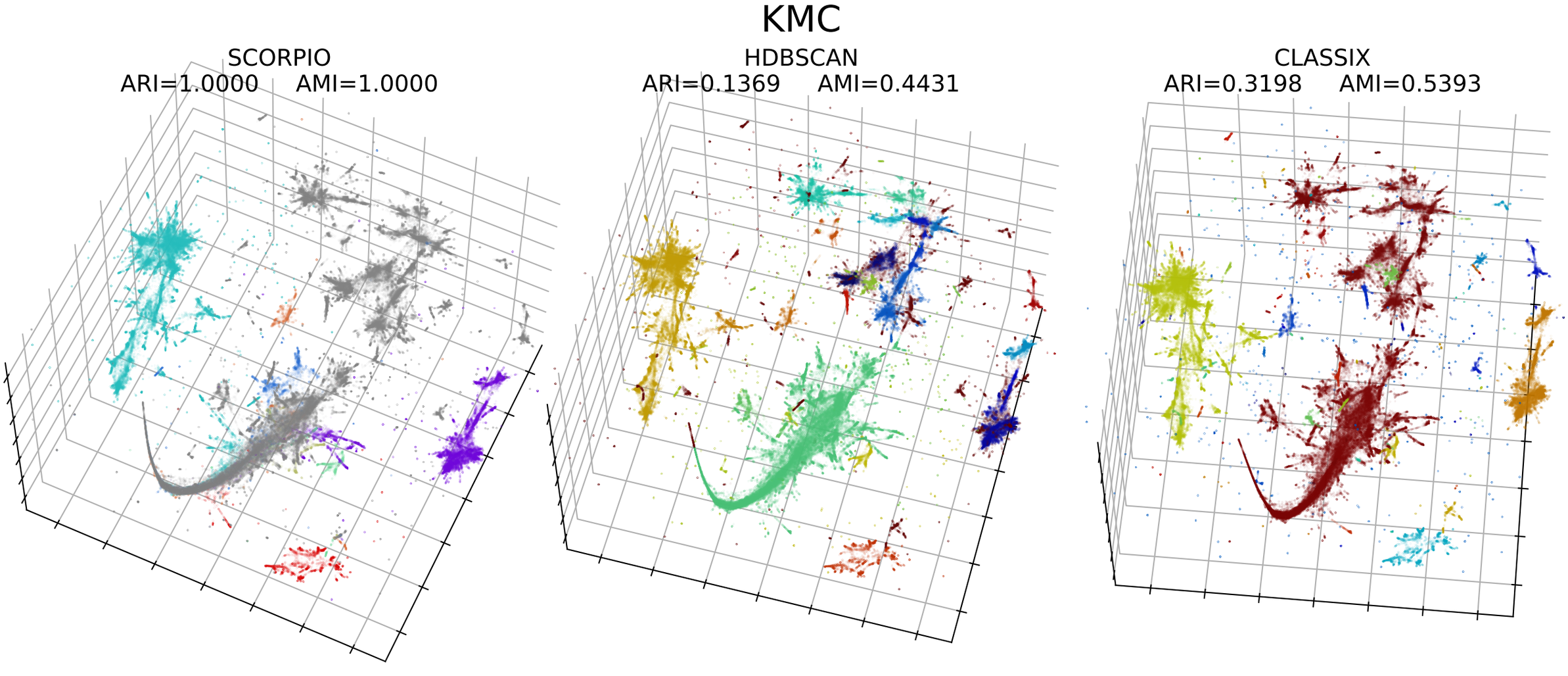
Figure S5: PaCMAP projection of the dataset extracting the WMRV features [16], (left) colouring based on Scorpio labelling [24], (centre) unsupervised clustering identification using HDBSCAN [19], (right) unsupervised clustering identification using CLASSIX [4]

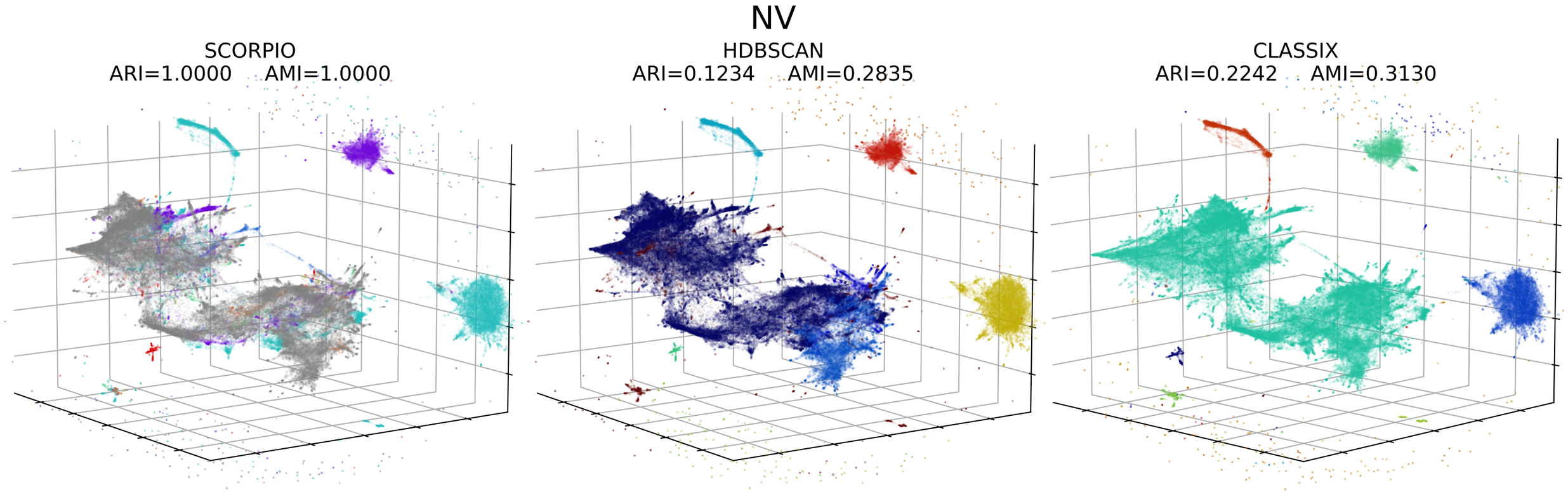
S.3 Supplementary Table

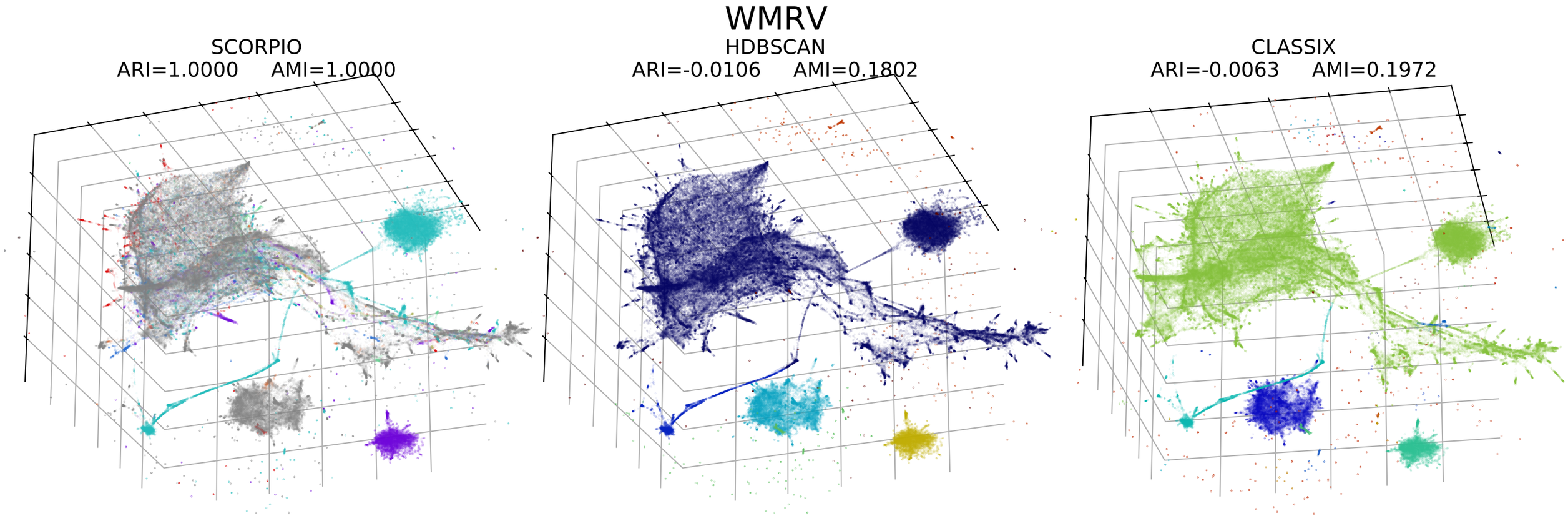
	HDBSCAN		**CLASSIX**	
Features	**ARI**	**AMI**	**ARI**	**AMI**
KMC	0.1369	0.4431	0.3198	0.5393
ENV	0.3906	0.4574	0.3837	0.4670
KMNV	0.0892	0.4032	0.2418	0.5010
NV	0.1234	0.2835	0.2242	0.3130
ANV	0.0590	0.2528	0.1608	0.2525
WMRV	-0.0106	0.1802	-0.0063	0.1972

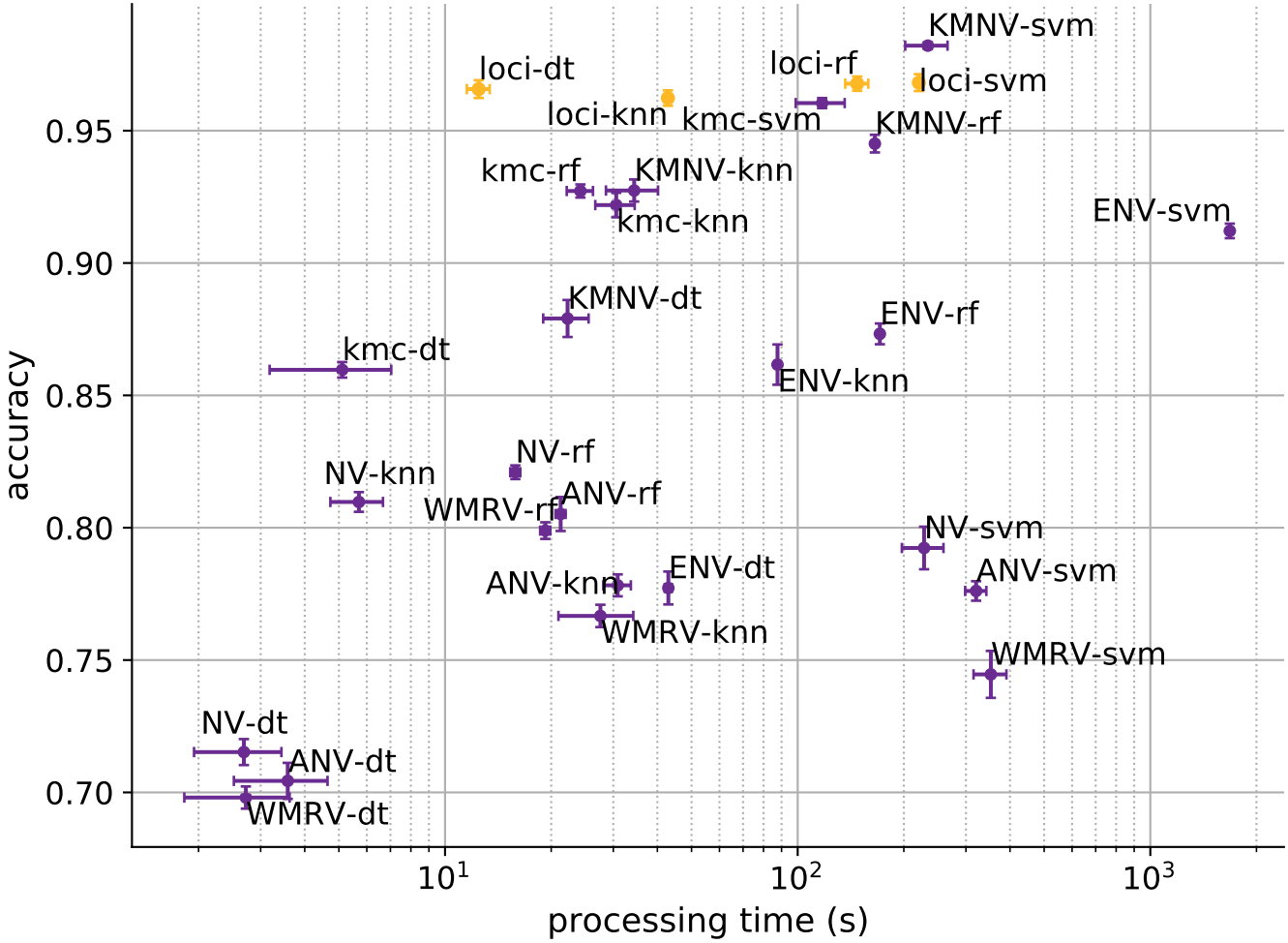
Table S1: Summary of ARI and AMI clustering evaluation for all word-statistics features

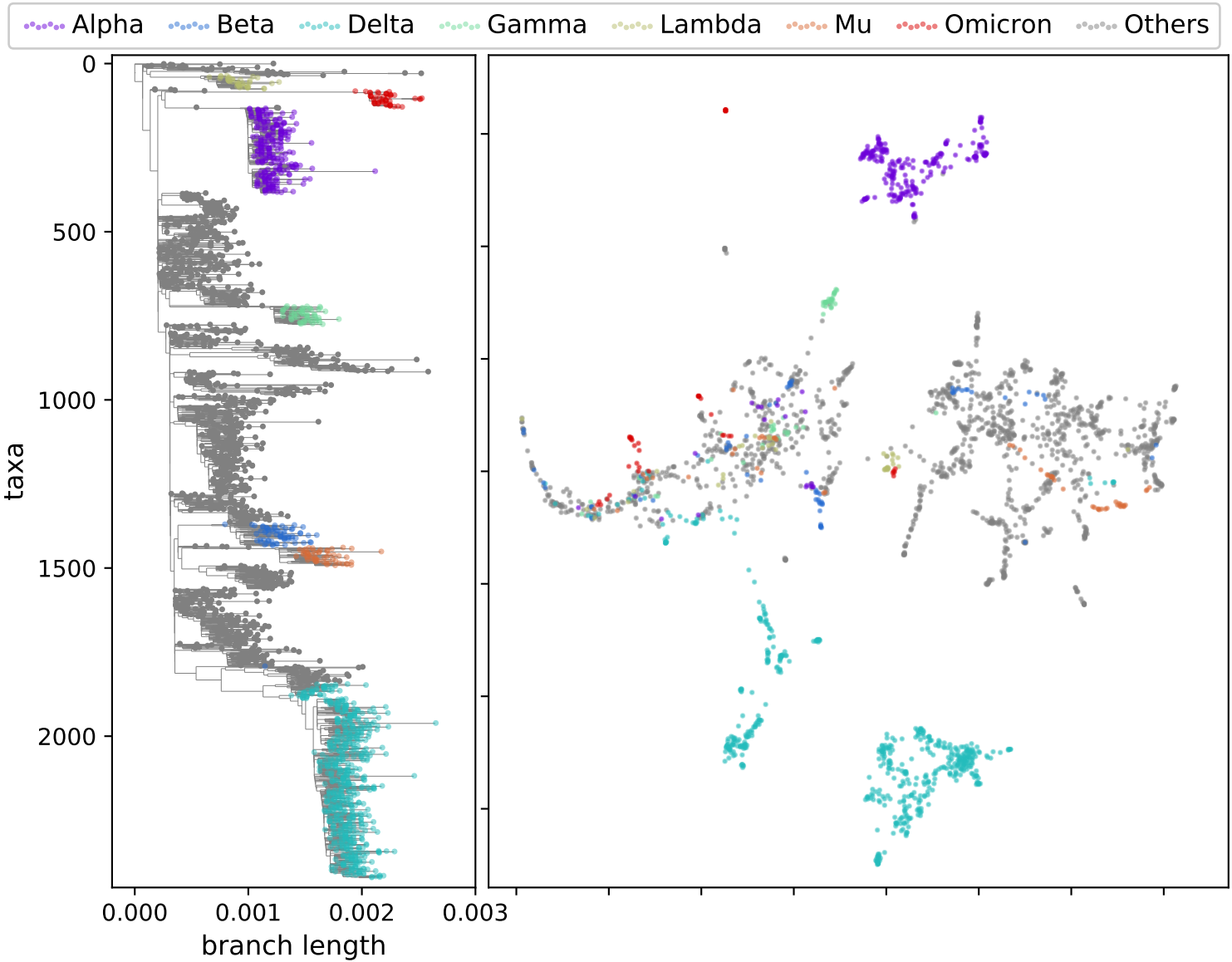

ENV
SCORPIO
ARI=1.0000 AMI=1.0000
HDBSCAN
ARI=0.3906 AMI=0.4574
CLASSIX
ARI=0.3837 AMI=0.4670

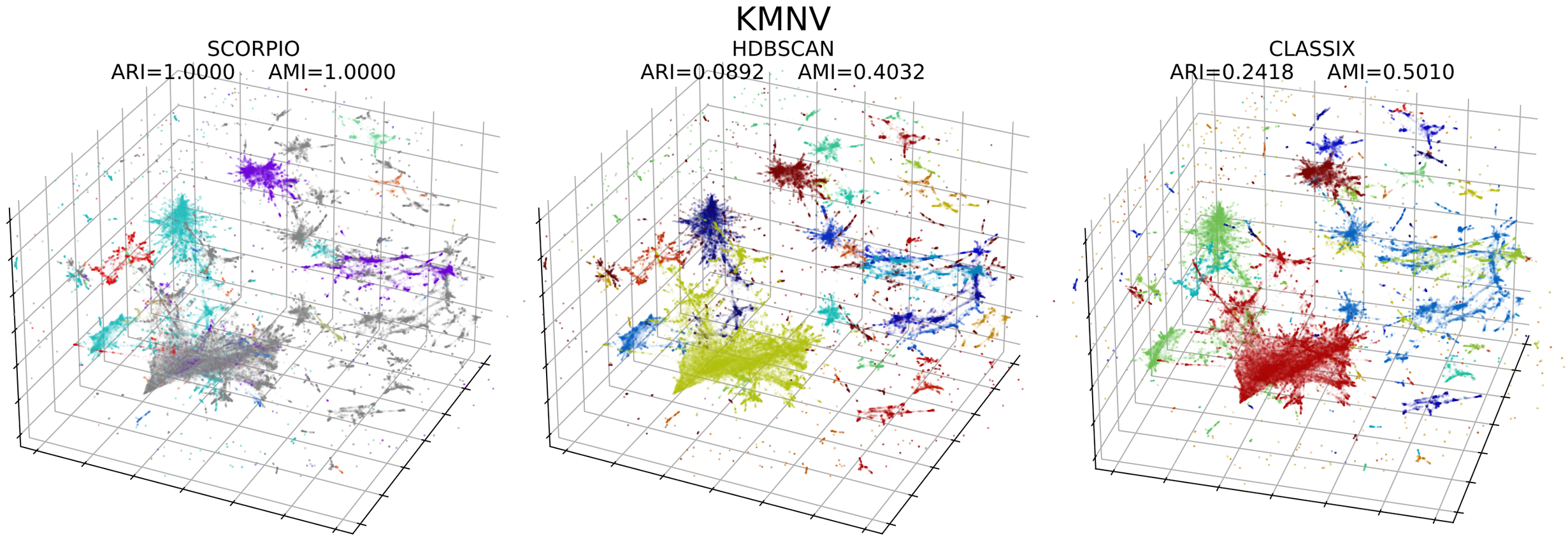

KMC
SCORPIO
ARI=1.0000 AMI=1.0000
HDBSCAN
ARI=0.1369 AMI=0.4431
CLASSIX
ARI=0.3198 AMI=0.5393

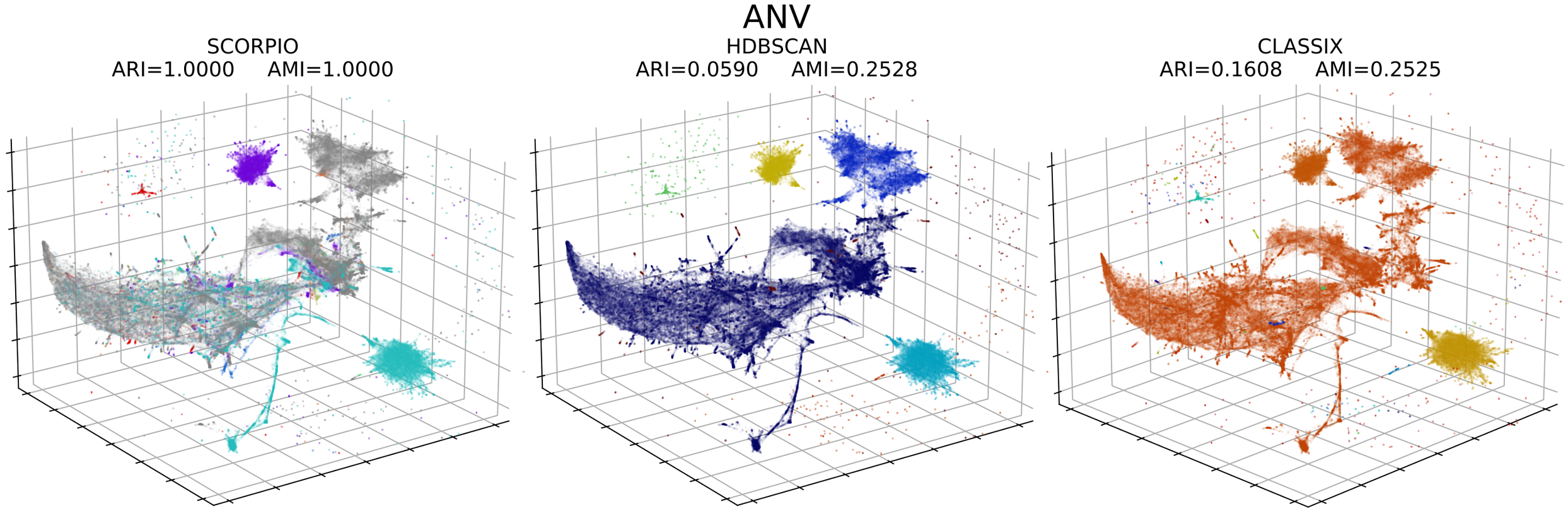

NV
SCORPIO
ARI=1.0000 AMI=1.0000
HDBSCAN
ARI=0.1234 AMI=0.2835
CLASSIX
ARI=0.2242 AMI=0.3130

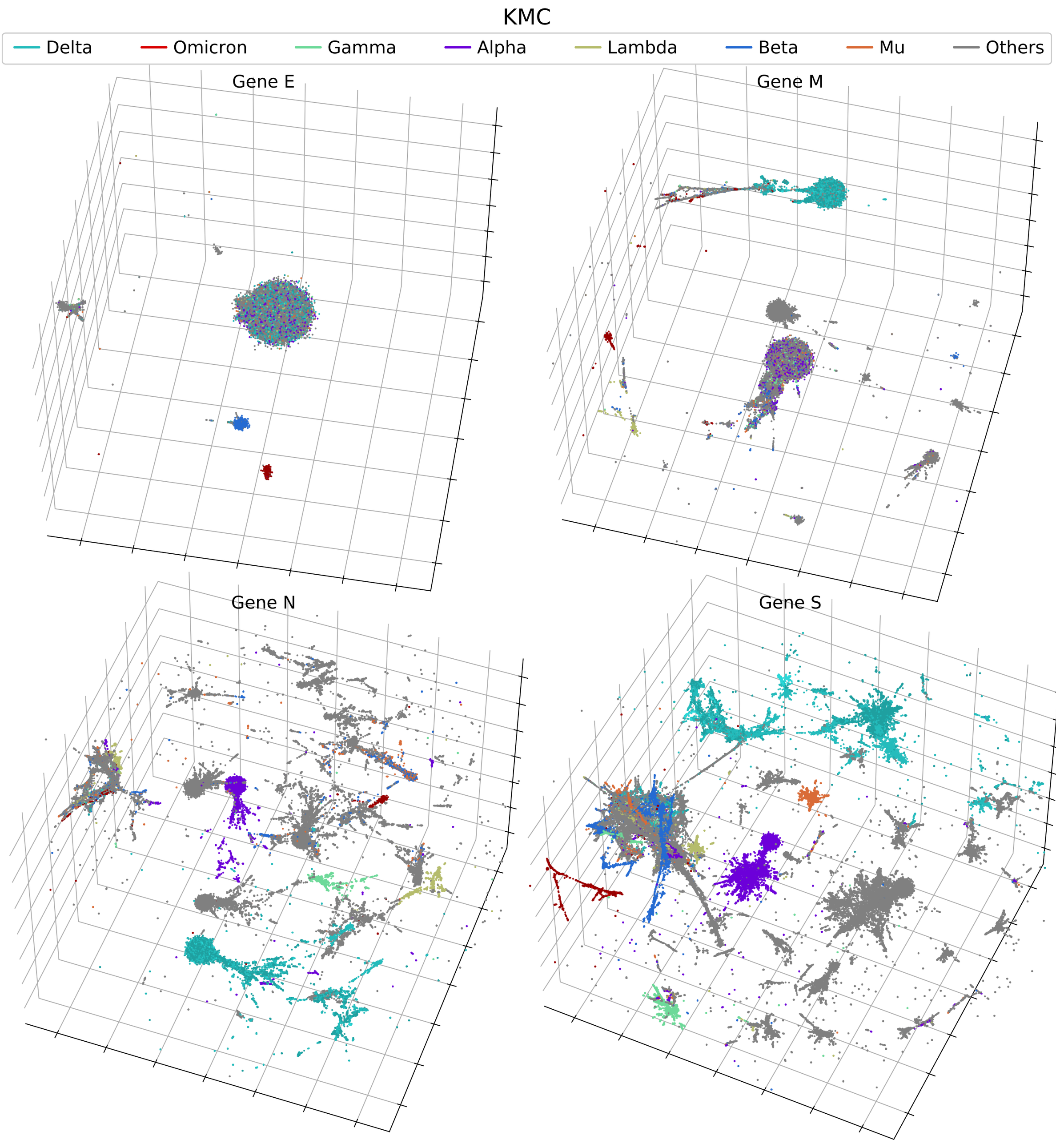

WMRV
SCORPIO
ARI=1.0000 AMI=1.0000
HDBSCAN
ARI=-0.0106 AMI=0.1802
CLASSIX
ARI=-0.0063 AMI=0.1972


KMNV-svm
loci-dt
loci-rf
loci-svm
loci-knn
kmc-svm
KMNV-rf
kmc-rf
KMNV-knn
kmc-knn
ENV-svm
KMNV-dt
ENV-rf
kmc-dt
ENV-knn
NV-rf
NV-knn
ANV-rf
WMRV-rf
NV-svm
ENV-dt
ANV-svm
ANV-knn
WMRV-knn
WMRV-svm
NV-dt
ANV-dt
WMRV-dt
0.95
0.90
0.85
0.80
0.75
0.70
accuracy
10^1
10^2
10^3
processing time (s)


Alpha
Beta
Delta
Gamma
Lambda
Mu
Omicron
Others
taxa
0
500
1000
1500
2000
0.000
0.001
0.002
0.003
branch length


KMNV
SCORPIO
ARI=1.0000 AMI=1.0000
HDBSCAN
ARI=0.0892 AMI=0.4032
CLASSIX
ARI=0.2418 AMI=0.5010


ANV
SCORPIO
ARI=1.0000 AMI=1.0000
HDBSCAN
ARI=0.0590 AMI=0.2528
CLASSIX
ARI=0.1608 AMI=0.2525


KMC
Delta
Omicron
Gamma
Alpha
Lambda
Beta
Mu
Others
Gene E
Gene M
Gene N
Gene S