Preserving Hidden Hierarchical Structure: Poincaré Distance for Enhanced Genomic Sequence Analysis

Sarwan Ali, Haris Mansoor, Prakash Chourasia, Imdadullah Khan, Murray Patterson

Georgia State University

December 9, 2024

Table of Contents

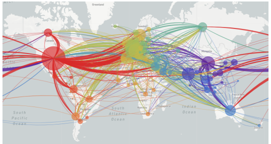
1. Motivation
2. Real World Application
3. Challenges
4. Proposed Model
5. Datasets
6. Data Visualization
7. Results
8. Conclusion

Motivation

- Studies of alterations in the protein sequence to classify and predict amino acid changes in SARS-CoV-2 are crucial in
  - Understanding the immune evasion and host-to-host transmission properties of SARS-CoV-2 and its variants
  - Identifying transmission patterns of each variant may help policymakers to prevent the rapid spread
  - Knowledge of mutations and variants will help identify transmission patterns
  - This will also help in vaccine design and efficacy
- Insights into the evolutionary relationships between organisms, helping us understand the origins and diversity of life on Earth.

Real World Application

- Genomic surveillance: Tracking the spread of pathogens in terms of genomic content
- Real time identification of new and rapidly emerging coronavirus variants
- Track the spread of known coronavirus variants in new municipalities, regions, countries and continents



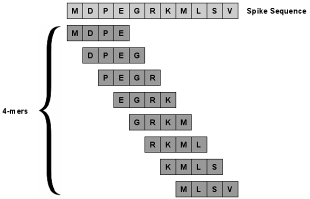
Challenges

- Mutations happen at different rates in different regions of the genome
- Since new variants (for coronavirus) are emerging, not much information is available about these variants
- The size of the data (millions of sequences) pose bottlenecks for traditional (e.g., phylogenetic) approaches
- Generating fixed-length feature vectors from variable-length sequences
- High dimensionality of generated embeddings (e.g., OHE)
- Challenges:
  - Preserving Hidden Hierarchical Structure from Sequences
  - Predictive Performance

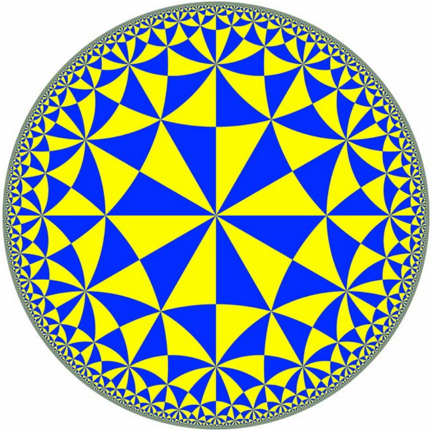
Feature Vector Representation

- To convert the sequences into fixed-length numerical representations, we use a recently proposed method called Spike2Vec [1].
- Spike2Vec generates a fixed-length numerical representation using the concept of k-mers (also called n-gram) for a sequence.

- It uses the idea of the sliding window to generate substrings (called mers) of length k (size of the window).
- From a set of k-mers from a sequence, a feature vector of length $|\Sigma|^k$ (Σ is the set of alphabets amino acid or nucleotide), is generated using the frequency/count of each k-mer.



Distance Computation (Poincaré distance)



$$d(x,y) = arcosh\left(1 + 2\frac{\|x-y\|^2}{(1-\|x\|^2)(1-\|y\|^2)}\right) \tag{1}$$

where

- $\|\cdot\|$ denotes the Euclidean norm
- arcosh(z) is the inverse hyperbolic cosine (cosh) function

Distance Computation (Modified-Poincaré distance)

- We propose a distance function (a modified form of Poincaré distance) that combines elements of Euclidean norms and dot products.
- Our distance function calculates the distance between two vectors x and y. This distance is a measure of dissimilarity and is defined as:

$$d'(x,y) = arcosh\left(\frac{1+\frac{2||x-y||^2}{(1-||x||^2)(1-||y||^2)}}{1-\frac{(x\cdot y)^2}{||x||^2||y||^2}}\right) \tag{2}$$

- where $d'(x,y)$ represents the modified Poincaré distance between vectors x and y , and $x \cdot y$ represents the dot product of vectors x and y.

Algorithm Poincaré Kernel Matrix

```
Input: Set of molecular sequences $S$
Output: Poincaré kernel matrix K
embed ← KMERSSPECTRUM(S)                                  ▷ Using method from [2]
K ← np.zeros($|S|$, $|S|$)
Initialize kernel matrix $K$ with zeros
for $i$ in range(len(embed)) do                           ▷ for all embeddings $i$
    for $j$ in range(len(embed)) do                       ▷ for all embeddings $j$
        if $i \leq j$ then
            Set $\sigma_{\text{val}} = 1$
            dist ← POINCAREDIST(embed[i], embed[j])       ▷ Eq. 1
            kVal = np.exp(-$\frac{power(dist,2)}{2 \times power(\sigma_{\text{val}},2)}$)   ▷ Gaussian Kernel
            Set $K[i,j] = kVal$
        else
            Set $K[j,i] = K[i,j]$
Return $K$
```

Dataset Statistics

Name	| Seq. |	| Classes |	Sequence Statistics			Reference	Description
			Max	Min	Mean		
Spike7k	7000	22	1274	1274	1274.00	[3]	The spike protein sequences of the SARS-CoV-2 virus having the information about the coronavirus Lineages of each sequence.
Human DNA	4380	7	18921	5	1263.59	[4]	Unaligned nucleotide sequences to classify gene family to which humans belong
Coronavirus Host	5558	21	1584	9	1272.36	ViPR [5], GISAID [3]	The spike protein sequences belonging to various clades of the Coronaviridae family accompanied by the infected host label e.g. Humans, Bats, Chickens, etc.

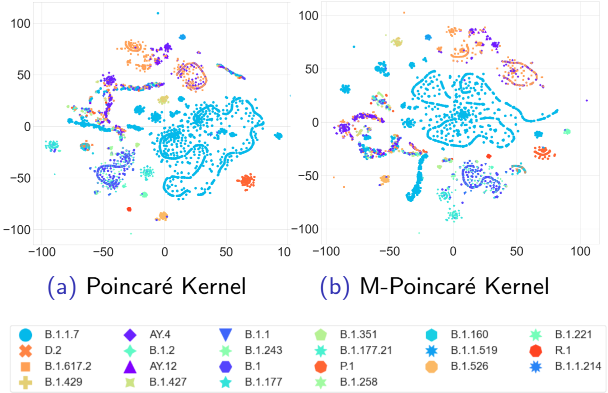
Table: Dataset Statistics for all three datasets that are used in performing the evaluation.

Baselines

Method	Category	Detail	Source
PWM2Vec	Feature Engineering	Take molecular sequence as input and design fixed-length numerical embeddings	[6]
String Kernel	Kernel Matrix	Designs $n \times n$ kernel matrix that can be used with kernel classifiers or with kernel PCA to get feature vector based on principal components	[7, 8]
WDGRL	Neural Network (NN)	Take one-hot representation of molecular sequence as input and design NN-based embedding method by minimizing loss	[9]
AutoEncoder			[10]
SeqVec	Pretrained Language Model	Takes molecular sequences as input and fine-tunes the weights based on a pre-trained model to get final embedding	[11]
ProteinBERT	Pretrained Transformer	A pre-trained protein sequence model to classify the given molecular sequence using Transformer/BERT	[12]

Table: Different baselines and SOTA methods description.

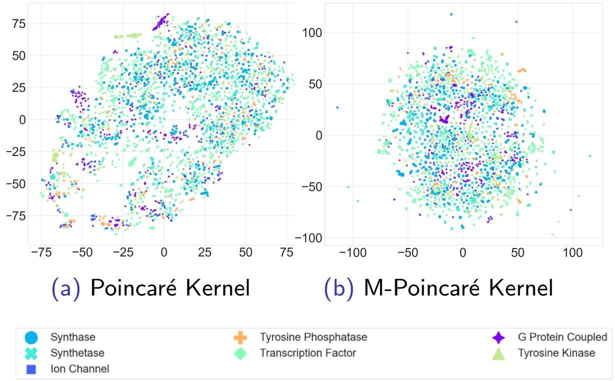
tSNE Results



(a) Poincaré Kernel (b) M-Poincaré Kernel

Figure: t-SNE plots for the proposed Poincaré and M-Poincaré kernels for the **Spike7k** dataset. These plots are generated after applying kernel PCA-based embeddings computed from both kernel methods. The legends show the lineages (target labels) for the Spike7k dataset.

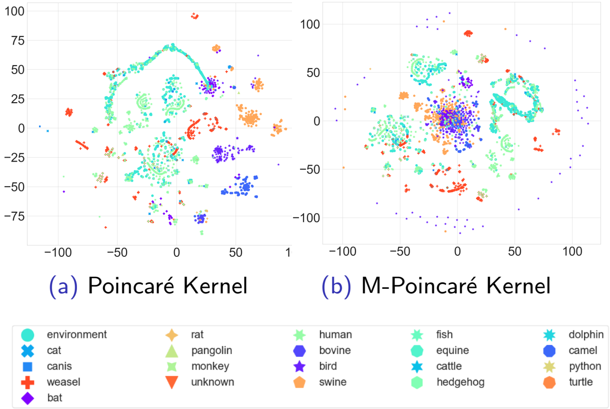
tSNE Results



(a) Poincaré Kernel (b) M-Poincaré Kernel

Figure: t-SNE plots for the proposed Poincaré and M-Poincaré kernels for the **Human DNA** dataset. The legends show the gene family (target labels) for the Human DNA dataset.

tSNE Results



(a) Poincaré Kernel

(b) M-Poincaré Kernel

Figure: t-SNE plots for the proposed Poincaré and M-Poincaré kernels for the **Coronavirus Host** dataset. The legends show the host names (target labels) for the Coronavirus Host dataset.

Classification Results - Spike7k Dataset

Embeddings	Algo.	Acc. ↑	Prec. ↑	Recall ↑	F1 (Weig.) ↑	F1 (Macro) ↑	ROC AUC ↑	Train Time (sec.) ↓
PWM2Vec	SVM	0.818	0.820	0.818	0.810	0.606	0.807	22.710
	NB	0.610	0.667	0.610	0.607	0.218	0.631	1.456
	MLP	0.812	0.792	0.812	0.794	0.530	0.770	35.197
	KNN	0.767	0.790	0.767	0.760	0.565	0.773	1.033
	RF	0.824	0.843	0.824	0.813	0.616	0.803	8.290
	LR	0.822	0.813	0.822	0.811	0.605	0.802	471.659
	DT	0.803	0.800	0.803	0.795	0.581	0.791	4.100
String Kernel	SVM	0.845	0.833	0.846	0.821	0.631	0.812	7.350
	NB	0.753	0.821	0.755	0.774	0.602	0.825	0.178
	MLP	0.831	0.829	0.838	0.823	0.624	0.818	12.652
	KNN	0.829	0.822	0.827	0.827	0.623	0.791	0.326
	RF	0.847	0.844	0.841	0.835	0.666	0.824	1.464
	LR	0.845	0.843	0.843	0.826	0.628	0.812	1.869
	DT	0.822	0.829	0.824	0.829	0.631	0.826	0.243
WDGRL	SVM	0.792	0.769	0.792	0.772	0.455	0.736	0.335
	NB	0.724	0.755	0.724	0.726	0.434	0.727	0.018
	MLP	0.799	0.779	0.799	0.784	0.505	0.755	7.348
	KNN	0.800	0.799	0.800	0.792	0.546	0.766	0.094
	RF	0.796	0.793	0.796	0.789	0.560	0.776	0.393
	LR	0.752	0.693	0.752	0.716	0.262	0.648	0.091
	DT	0.790	0.799	0.790	0.788	0.557	0.768	**0.009**
Auto-Encoder	SVM	0.699	0.720	0.699	0.678	0.243	0.627	4018.028
	NB	0.490	0.533	0.490	0.481	0.123	0.620	24.6372
	MLP	0.663	0.633	0.663	0.632	0.161	0.589	87.4913
	KNN	0.782	0.791	0.782	0.776	0.535	0.761	24.5597
	RF	0.814	0.803	0.814	0.802	0.593	0.793	46.583
	LR	0.761	0.755	0.761	0.735	0.408	0.705	11769.02
	DT	0.803	0.792	0.803	0.792	0.546	0.779	102.185
SeqVec	SVM	0.796	0.768	0.796	0.770	0.479	0.747	1.0996
	NB	0.686	0.703	0.686	0.686	0.351	0.694	0.0146
	MLP	0.796	0.771	0.796	0.771	0.510	0.762	13.172
	KNN	0.790	0.787	0.790	0.786	0.561	0.768	0.6463
	RF	0.793	0.788	0.793	0.786	0.557	0.769	1.8241
	LR	0.785	0.763	0.785	0.761	0.459	0.740	1.7535
	DT	0.757	0.756	0.757	0.755	0.521	0.760	0.1308
Protein Bert	-	0.836	0.828	0.836	0.814	0.570	0.792	14163.52
Poincaré (ours)	SVM	0.484	0.235	0.484	0.316	0.030	0.500	5.789
	NB	0.215	0.663	0.215	0.213	0.357	0.703	0.149
	MLP	0.740	0.734	0.740	0.731	0.526	0.760	18.037
	KNN	0.808	0.812	0.808	0.805	0.630	0.803	0.512
	RF	0.798	0.794	0.798	0.780	0.629	0.789	10.054
	LR	0.484	0.235	0.484	0.316	0.030	0.500	3.828
	DT	0.804	0.803	0.804	0.799	0.623	**0.833**	1.581
M-Poincaré (ours)	SVM	0.605	0.457	0.605	0.490	0.090	0.532	7.592
	NB	0.187	0.442	0.187	0.225	0.223	0.645	0.440
	MLP	0.713	0.724	0.713	0.707	0.478	0.738	5.155
	KNN	0.816	0.821	0.816	0.811	0.620	0.798	0.225
	RF	**0.851**	**0.849**	**0.851**	**0.840**	**0.669**	0.796	5.622
	LR	0.476	0.248	0.476	0.326	0.029	0.498	3.647
	DT	0.787	0.797	0.787	0.786	0.584	0.777	1.517

Classification Results - Human DNA Dataset

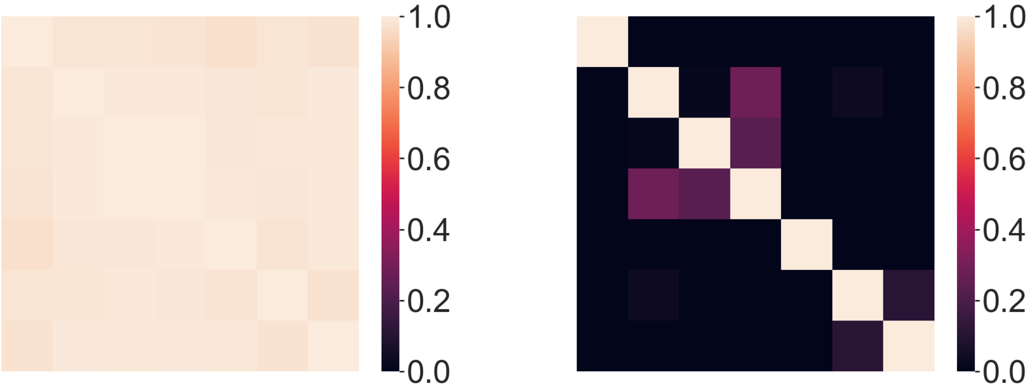
Embeddings	Algo.	Acc. ↑	Prec. ↑	Recall ↑	F1 (Weig.) ↑	F1 (Macro) ↑	ROC AUC ↑	Train Time (sec.) ↓
PWM2Vec	SVM	0.302	0.241	0.302	0.165	0.091	0.505	10011.3
	NB	0.084	0.442	0.084	0.063	0.066	0.511	4.565
	MLP	0.310	0.350	0.310	0.175	0.107	0.510	320.555
	KNN	0.121	0.337	0.121	0.093	0.077	0.509	2.193
	RF	0.309	0.332	0.309	0.181	0.110	0.510	65.250
	LR	0.304	0.257	0.304	0.167	0.094	0.506	23.651
	DT	0.306	0.284	0.306	0.181	0.111	0.509	1.861
String Kernel	SVM	0.618	0.617	0.618	0.613	0.588	0.753	39.791
	NB	0.338	0.452	0.338	0.347	0.333	0.617	0.276
	MLP	0.597	0.595	0.597	0.593	0.549	0.737	331.068
	KNN	0.645	0.657	0.645	0.646	0.612	0.774	1.274
	RF	0.731	0.776	0.731	0.729	0.723	0.808	12.673
	LR	0.571	0.570	0.571	0.558	0.532	0.716	2.995
	DT	0.630	0.631	0.630	0.630	0.598	0.767	2.682
WDGRL	SVM	0.318	0.101	0.318	0.154	0.069	0.500	0.751
	NB	0.232	0.214	0.232	0.196	0.138	0.517	**0.004**
	MLP	0.326	0.286	0.326	0.263	0.186	0.535	8.613
	KNN	0.317	0.317	0.317	0.315	0.266	0.574	0.092
	RF	0.453	0.501	0.453	0.430	0.389	0.625	1.124
	LR	0.323	0.279	0.323	0.177	0.095	0.507	0.041
	DT	0.368	0.372	0.368	0.369	0.328	0.610	0.047
Auto-Encoder	SVM	0.621	0.638	0.621	0.624	0.593	0.769	22.230
	NB	0.260	0.426	0.260	0.247	0.268	0.583	0.287
	MLP	0.621	0.624	0.621	0.620	0.578	0.756	111.809
	KNN	0.565	0.577	0.565	0.568	0.547	0.732	1.208
	RF	0.689	0.738	0.689	0.683	0.668	0.774	20.131
	LR	0.692	0.700	0.692	0.693	0.672	0.799	58.369
	DT	0.543	0.546	0.543	0.543	0.515	0.718	10.616
SeqVec	SVM	0.656	0.661	0.656	0.652	0.611	0.791	0.891
	NB	0.324	0.445	0.312	0.295	0.282	0.624	0.036
	MLP	0.657	0.633	0.653	0.646	0.616	0.783	12.432
	KNN	0.592	0.606	0.592	0.591	0.552	0.717	0.571
	RF	0.713	0.724	0.701	0.702	0.693	0.752	2.164
	LR	0.725	0.715	0.726	0.725	0.685	0.784	1.209
	DT	0.586	0.553	0.585	0.577	0.557	0.736	0.24
Protein Bert	-	0.542	0.580	0.542	0.514	0.447	0.675	58681.57
Poincaré Kernel (ours)	SVM	0.307	0.094	0.307	0.144	0.067	0.500	10.709
	NB	0.149	0.345	0.149	0.114	0.114	0.522	0.086
	MLP	0.660	0.660	0.660	0.659	0.616	0.779	28.152
	KNN	0.647	0.660	0.647	0.650	0.611	0.774	0.540
	RF	**0.764**	0.792	**0.764**	**0.762**	**0.756**	**0.832**	12.927
	LR	0.307	0.094	0.307	0.144	0.067	0.500	2.009
	DT	0.608	0.617	0.608	0.611	0.574	0.758	4.373
M-Poincaré Kernel (ours)	SVM	0.353	0.471	0.353	0.223	0.136	0.525	12.650
	NB	0.309	0.434	0.309	0.306	0.295	0.596	0.131
	MLP	0.677	0.684	0.677	0.678	0.656	0.804	22.044
	KNN	0.714	0.728	0.714	0.716	0.685	0.827	0.515
	RF	0.743	**0.817**	0.743	0.745	0.747	0.812	20.750
	LR	0.374	0.482	0.374	0.272	0.186	0.541	3.366
	DT	0.585	0.590	0.585	0.586	0.558	0.746	9.763

Classification Results - Coronavirus Host Dataset

Embeddings	Algo.	Acc.	Prec.	Recall	F1 (Weig.)	F1 (Macro)	ROC AUC	Train Time (Sec.)
PWM2Vec	SVM	0.799	0.806	0.799	0.801	0.648	0.859	44.793
	NB	0.381	0.584	0.381	0.358	0.400	0.683	2.494
	MLP	0.782	0.792	0.782	0.778	0.693	0.848	21.191
	KNN	0.786	0.782	0.786	0.779	0.679	0.838	12.933
	RF	0.836	0.839	0.836	0.828	**0.739**	0.862	7.690
	LR	0.809	0.815	0.809	0.800	0.728	0.852	274.917
	DT	0.801	0.802	0.801	0.797	0.633	0.829	4.537
String Kernel	SVM	0.601	0.673	0.601	0.602	0.325	0.624	5.198
	NB	0.230	0.665	0.230	0.295	0.162	0.625	0.131
	MLP	0.647	0.696	0.647	0.641	0.302	0.628	42.322
	KNN	0.613	0.623	0.613	0.612	0.310	0.629	0.434
	RF	0.668	0.692	0.668	0.663	0.360	0.658	4.541
	LR	0.554	0.724	0.554	0.505	0.193	0.568	5.096
	DT	0.646	0.674	0.646	0.643	0.345	0.653	1.561
WDGRL	SVM	0.329	0.108	0.329	0.163	0.029	0.500	2.859
	NB	0.004	0.095	0.004	0.007	0.002	0.496	**0.008**
	MLP	0.328	0.136	0.328	0.170	0.032	0.499	5.905
	KNN	0.235	0.198	0.235	0.211	0.058	0.499	0.081
	RF	0.261	0.196	0.261	0.216	0.051	0.499	1.288
	LR	0.332	0.149	0.332	0.177	0.034	0.500	0.365
	DT	0.237	0.202	0.237	0.211	0.054	0.498	0.026
Auto-Encoder	SVM	0.602	0.588	0.602	0.590	0.519	0.759	2575.955
	NB	0.261	0.520	0.261	0.303	0.294	0.673	21.7474
	MLP	0.486	0.459	0.486	0.458	0.216	0.594	29.93393
	KNN	0.763	0.764	0.763	0.755	0.547	0.784	18.51143
	RF	0.800	0.796	0.800	0.791	0.648	0.815	57.90582
	LR	0.717	0.750	0.717	0.702	0.564	0.812	11072.67
	DT	0.772	0.767	0.772	0.765	0.571	0.808	121.3628
SeqVec	SVM	0.711	0.745	0.711	0.698	0.497	0.747	0.751
	NB	0.503	0.636	0.503	0.554	0.413	0.648	0.012
	MLP	0.718	0.748	0.718	0.708	0.407	0.706	10.191
	KNN	0.815	0.806	0.815	0.809	0.588	0.800	0.418
	RF	0.833	0.824	0.833	0.828	0.678	0.839	1.753
	LR	0.673	0.683	0.673	0.654	0.332	0.660	1.177
	DT	0.778	0.786	0.778	0.781	0.618	0.825	0.160
Protein Bert	-	0.799	0.806	0.799	0.789	0.715	0.841	15742.95
Poincaré (ours)	SVM	0.334	0.115	0.334	0.169	0.056	0.510	35.848
	NB	0.594	0.694	0.594	0.579	0.461	0.749	0.729
	MLP	0.752	0.750	0.752	0.744	0.463	0.733	62.323
	KNN	0.793	0.789	0.793	0.789	0.645	0.815	1.543
	RF	**0.844**	**0.847**	**0.844**	**0.836**	0.687	**0.868**	35.333
	LR	0.333	0.111	0.333	0.167	0.028	0.500	18.710
	DT	0.795	0.794	0.795	0.791	0.546	0.786	10.869
M-Poincaré (ours)	SVM	0.332	0.195	0.332	0.175	0.032	0.501	25.668
	NB	0.450	0.524	0.450	0.424	0.296	0.642	0.351
	MLP	0.607	0.599	0.607	0.598	0.268	0.627	40.005
	KNN	0.678	0.710	0.678	0.684	0.353	0.686	0.912
	RF	0.788	0.792	0.788	0.778	0.480	0.714	19.283
	LR	0.346	0.242	0.346	0.212	0.043	0.504	10.930
	DT	0.733	0.734	0.733	0.730	0.417	0.721	6.422

Table: Classification results (averaged over 5 runs) for different evaluation metrics for **Coronavirus**

Heatmap Results

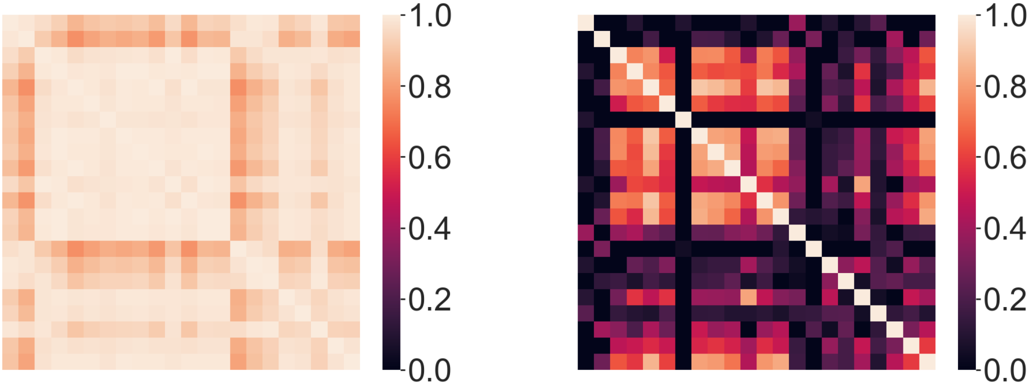


(a) *k*-mers spectrum

(b) M-Poincaré Kernel

Figure: Heatmap for classes in **Human DNA** dataset.

Heatmap Results

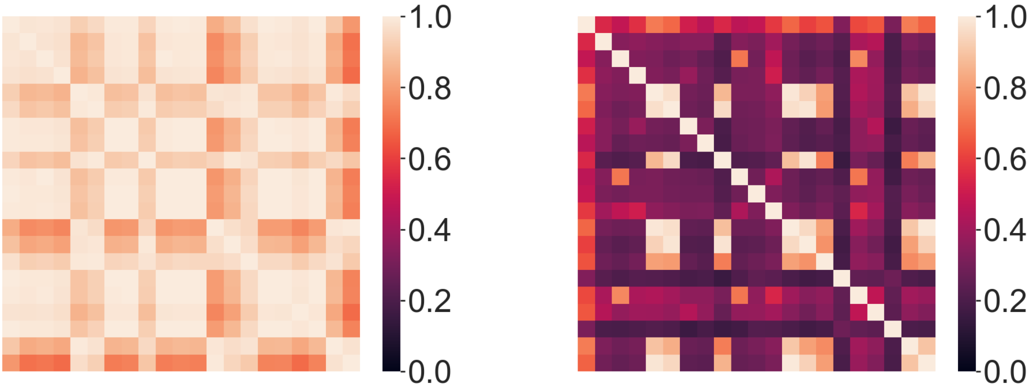


(a) *k*-mers spectrum

(b) M-Poincaré Kernel

Figure: Heatmap for classes in **Spike7K** dataset.

Heatmap Results



(a) *k*-mers spectrum

(b) M-Poincaré Kernel

Figure: Heatmap for classes in **Coronavirus Host** dataset.

Conclusion and Future Work

Conclusion

- we have addressed the limitations of traditional Euclidean-based distance measurements and discussed the concept of hyperbolic geometry, and proposed the use of Poincaré distance as a more effective and meaningful measure.
- By leveraging the unique properties of hyperbolic space, the Poincaré distance preserves the hierarchical structures present in molecular sequences.
- Furthermore, we introduced a modified version of the Poincaré distance, known as M-Poincaré, which combines Euclidean norms and the dot product between sequence representations.

Future Work

- Future research can explore the application of these methods in other domains along with interpretability studies.

Thank You

References

S. Ali and M. Patterson, “Spike2vec: An efficient and scalable embedding approach for covid-19 spike sequences,” in *IEEE International Conference on Big Data (Big Data)*, 2021, pp. 1533–1540.

S. Ali, B. Sahoo, N. Ullah, A. Zelikovskiy, M. Patterson, and I. Khan, “A k-mer based approach for SARS-CoV-2 variant identification,” in *International Symposium on Bioinformatics Research and Applications*, 2021, pp. 153–164.

GISAID Website, https://www.gisaid.org/, 2022, [Online; accessed 17-December-2022].

Human DNA, in *https://www.kaggle.com/code/nageshsingh/demystify-dna-sequencing-with-machine-learning/data*, 2022, [Online; accessed 10-October-2022].

B. E. Pickett, E. L. Sadat, Y. Zhang, J. M. Noronha, R. B. Squires, V. Hunt, M. Liu, S. Kumar, S. Zaremba, Z. Gu *et al.*, “Vipr: an open bioinformatics database and analysis resource for virology research,” *Nucleic acids research*, vol. 40, no. D1, pp. D593–D598,