Pymeris: A Re-implementation of Kameris for fast and accurate HIV-1 Subtyping

Reproducing Solis-Reyes et al. (2018) with a pure-Python stack (NumPy/SciPy/scikit-learn) to validate accuracy and workflow [1].

Experiment 1: Using the HIV-1 whole genome dataset, train 15 classifiers to identify the best performer.
Experiment 2 - 5: Testing the SVM classifier on other genome datasets for HIV-1's Pol gene, and other viruses like Dengue, Hepatitis B, Hepatitis C and Influenza A.
Experiment 6: Model evaluation test where the SVM model was trained on a subset of the HIV-1 pol genome and tested on a benchmark dataset built from a combination of other subtypes of the HIV-1 Genome.
Experiment 7: Testing the model's ability to differentiate synthetic HIV-1 sequences from Natural ones.
Experiment 8: Unsupervised visualization of subtype clusters as well as a Synthetic vs Natural sequences comparison.

GGTCTCTCNNGTTAGACCA
GATTTGAGCCTGGNAGCTC
TCTGGCTAACTAGGGAC…

Preprocessing: Feature Vector ($F_k(s)$)

K = 6, Count occurrence of 4^6 k-mers
Normalize by dividing by the sequence length

Scale $F_k(s)$ k-mer frequencies to standard deviation = 1
Truncated singular value decomposition to reduce dimensions to 10% of the avg non-zero entries in $F_k(s)$

Supervised Model Training

Used 15 of Scikit-Learn's classifier implementations
Mostly default settings and hyperparameters

10 - Fold Cross Validation

Cross-validation splits labeled data into multiple train/validation folds, training on one subset and evaluating on the held-out subset in each round. Performance is then quantified by averaging the result metric across all the training/validation rounds.

For Model training, subtype classes with total counts ≤ 18 were removed from their datasets to help with class balancing.

Linear SVM

Unsupervised Visualizations

Construct distance matrix using Manhattan distance

$$d_M(\boldsymbol{A},\boldsymbol{B}) = \sum_{i=1}^{n} |a_i - b_i|$$

Visualize with Multidimensional Scaling MoDMap

Table 1. Classifier performance on the Whole Genome dataset, Pymeris vs Kameris

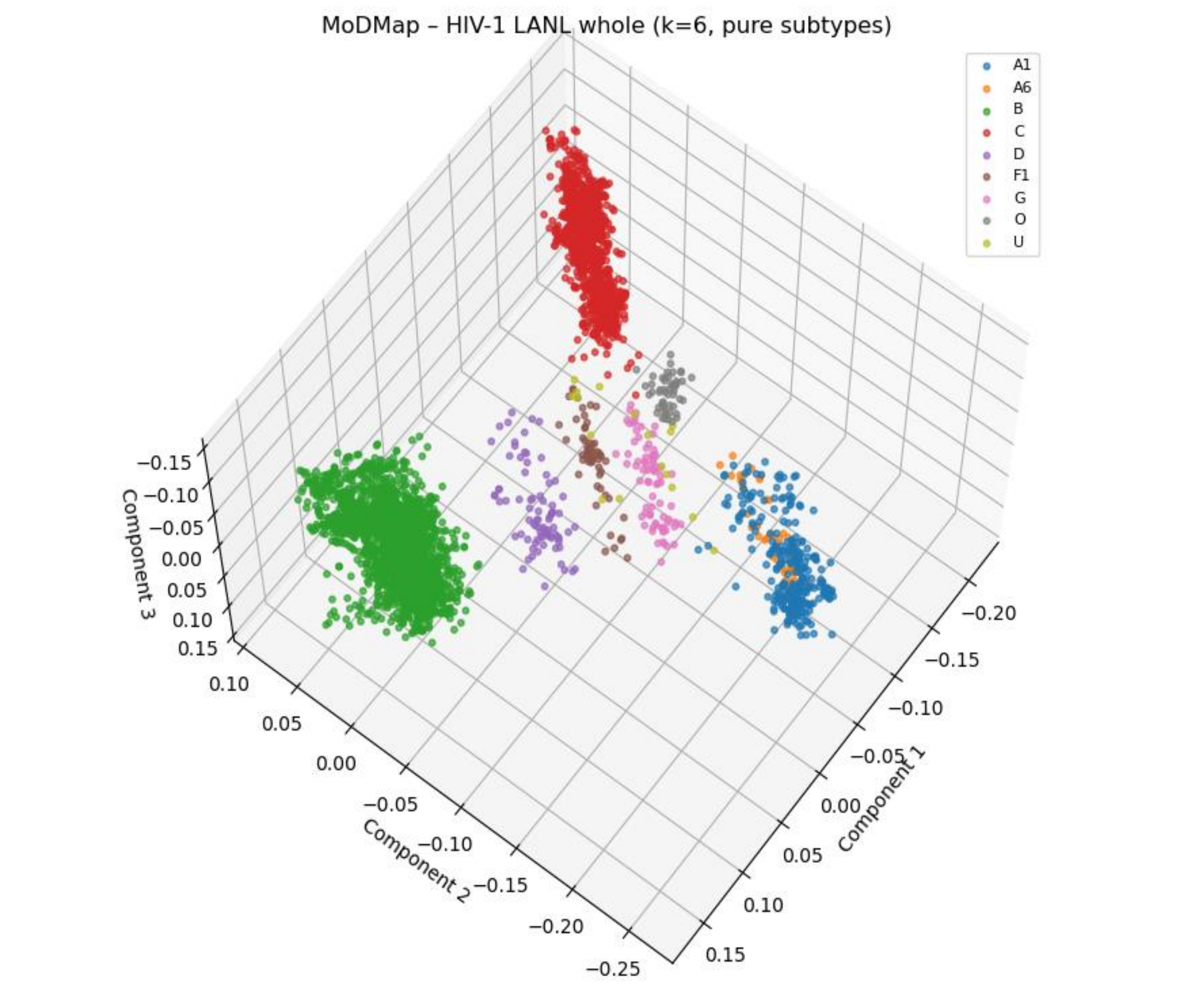
Model	Pymeris		Kameris	
	Mean Acc (%)	Mean Runtime (s)	Mean Acc (%)	Runtime (s)
Linear SVM	96.73	504.49	96.49	57.7
Logistic Regression	95.77	528.44	95.32	102.0
Multilayer Perceptron	95.69	621.97	95.49	60.6
LDA	95.09	486.27	77.76	36.0
Nearest Centroid (median)	94.34	481.16	93.95	34.0
10-Nearest Neighbors	94.07	490.24	93.97	44.3
Nearest Centroid (mean)	94.05	478.93	93.84	33.7
Decision Tree	93.75	502.90	93.53	62.3
AdaBoost	93.54	503.10	64.85	159.3
Cubic SVM	93.53	501.59	96.66	59.7
Random Forest	93.30	488.69	93.07	43.7
Quadratic SVM	92.85	504.36	96.59	58.3
SGD	88.84	481.20	91.10	37.4
Gaussian NB	88.49	478.67	87.75	34.0
QDA	75.51	490.52	75.13	38.3

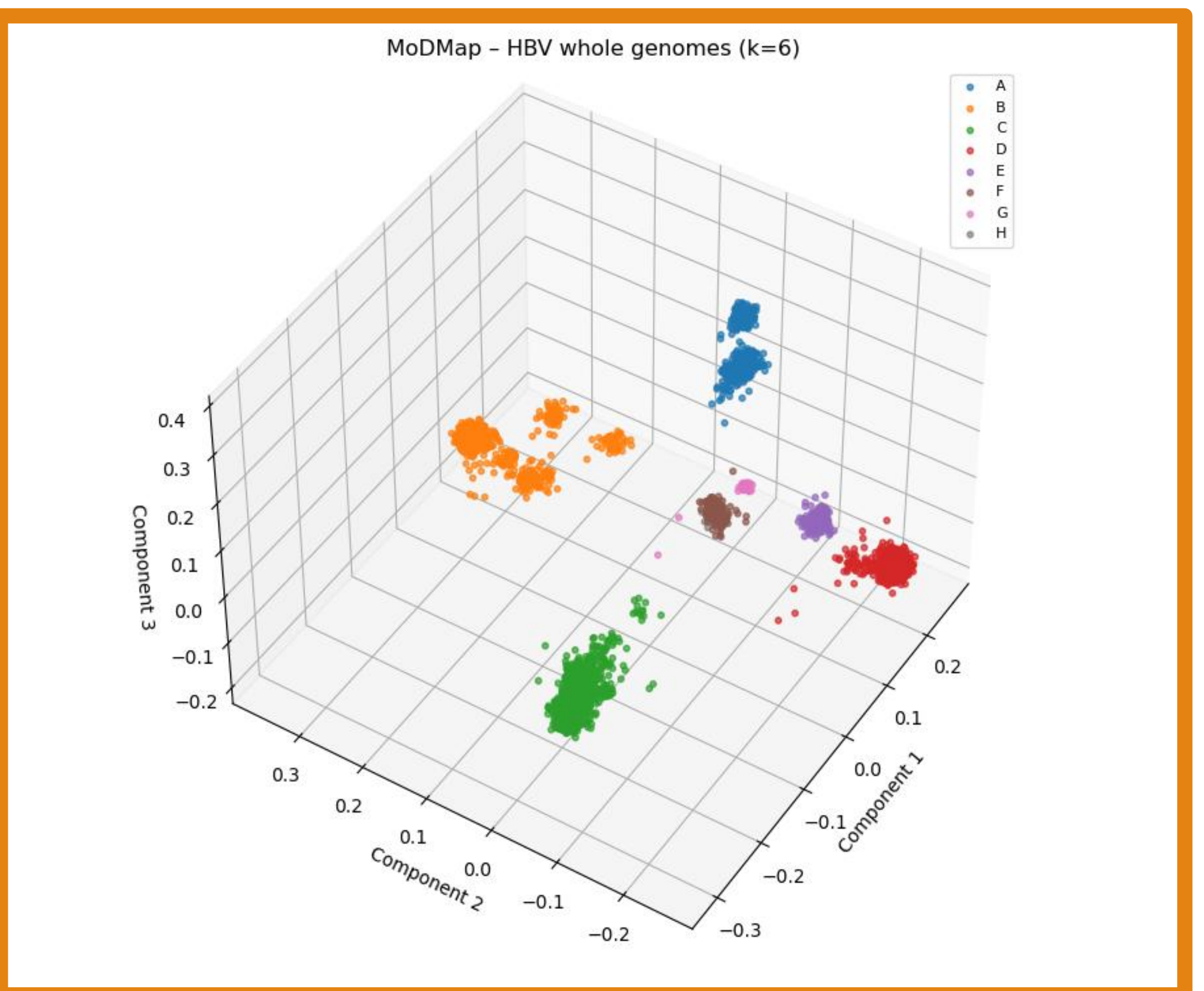
Table 2. Generalization experiments performance comparisons

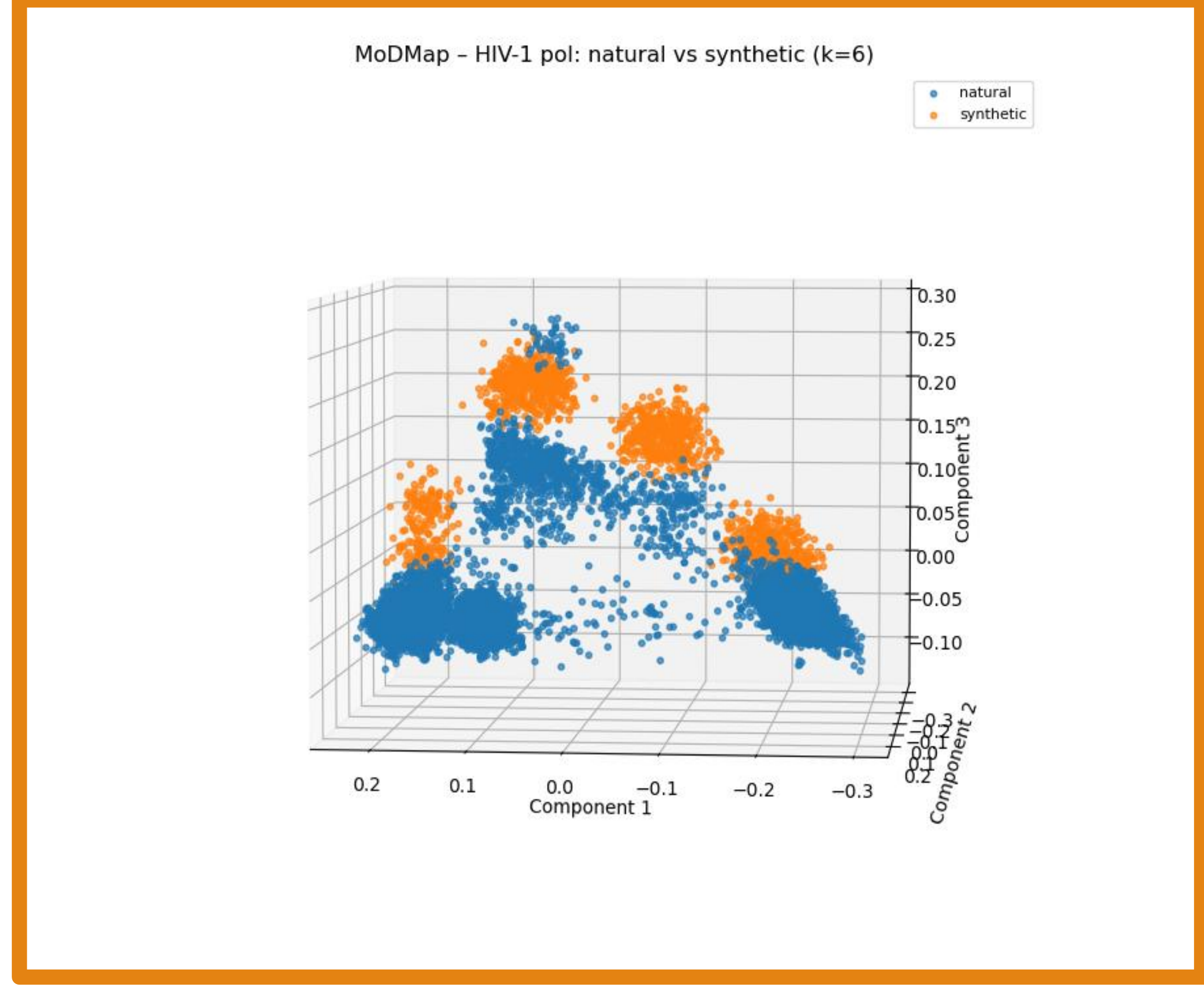
Experiment	Pymeris Accuracy	Kameris Accuracy
Hiv1 lanl pol	95.0%	95.68%
Dengue	99.98%	100%
Hepatitis B	94.96%	95.81%
Hepatitis C	99.94%	100%
Influenza A	96.65%	96.68%

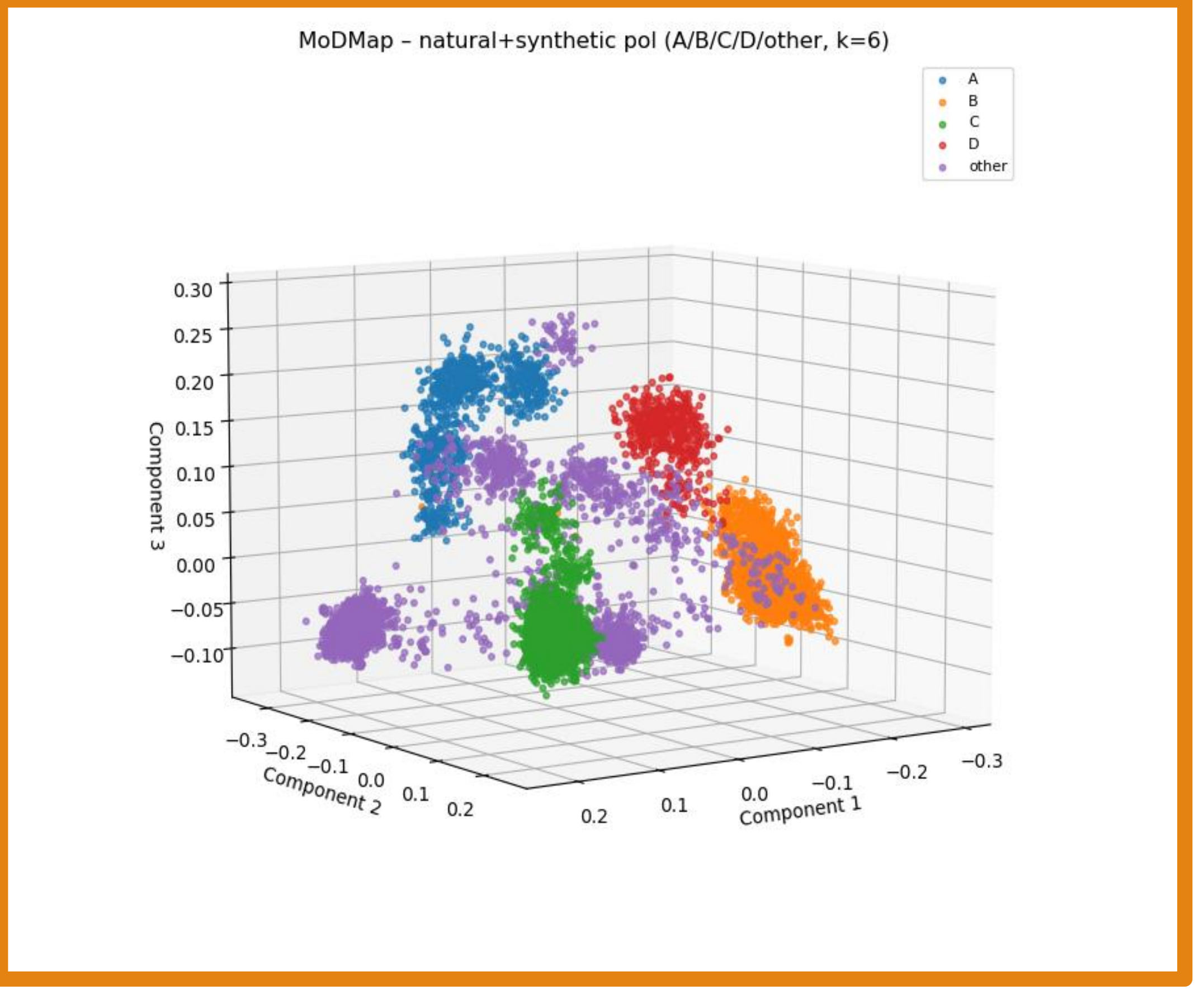
Table 3. Benchmark and Synthetic vs Natural genome comparisons

Experiment	Pymeris Accuracy	Kameris Accuracy
Benchmark test (reimplementation)	82.40%	94.3%
Synthetic vs natural pol fragments	99.98%	100%


MoDMap – HIV-1 LANL whole (k=6, pure subtypes)


MoDMap – HBV whole genomes (k=6)


MoDMap – HIV-1 pol: natural vs synthetic (k=6)


MoDMap – natural+synthetic pol (A/B/C/D/other, k=6)

[1] Solis-Reyes, S., Avino, M., Poon, A. and Kari, L. (2018) An open-source k-mer based machine learning tool for fast and accurate subtyping of HIV-1 genomes. PLOS ONE, 13(11), e0206409.