Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

GREAC: An Open-source software for Genome Region Extraction and Classification

Felipe Bueno de Souza
Universidade Tecnológica Federal do Paraná (UTFPR)
Matheus Henrique Pimenta-Zanon
Universidade Tecnológica Federal do Paraná (UTFPR)
Dora Henriques
CIMO, LA SusTEC, Instituto Politécnico de Bragança
Carlos Balsa
Research Centre in Digitalization and Intelligent Robotics
M. Alice Pinto
CIMO, LA SusTEC, Instituto Politécnico de Bragança
José Rufino
Research Centre in Digitalization and Intelligent Robotics
Fabricio Martins Lopes
fabricio@utfpr.edu.br

Universidade Tecnológica Federal do Paraná (UTFPR)

Method Article

Keywords: Bioinformatics, pattern recognition, viral genomics, K-mer analysis, discriminative regions, alignment-free analysis, data-driven, dimensionality reduction

Posted Date: July 7th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10241989/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

GREAC: An Open-source software for Genome Region Extraction and Classification

Felipe Bueno de Souza[1,2,3*], Matheus Henrique Pimenta-Zanon[1], Dora Henriques[3], Carlos Balsa[2], M. Alice Pinto[3], José Rufino[2], Fabricio Martins Lopes[1]

[1]Campus Cornélio Procópio, Universidade Tecnológica Federal do Paraná (UTFPR), Cornélio Procópio, PR, Brasil.
[2]CeDRI, SusTEC, Instituto Politécnico de Bragança, Campus de Santa Apolónia, 5300-253, Bragança, Portugal.
[3]CIMO, LA SusTEC, Instituto Politécnico de Bragança, Campus de Santa Apolónia, 5300-253, Bragança, Portugal.

*Corresponding author(s). E-mail(s): felipebuenosouza@alunos.utfpr.edu.br;
Contributing authors: omatheuspimenta@outlook.com; dorasmh@ipb.pt; balsa@ipb.pt; apinto@ipb.pt; rufino@ipb.pt; fabricio@utfpr.edu.br;

Abstract

Background: Advances in sequencing technologies open opportunities for research into genetic variants, disease associations, and the identification of genomic signatures of natural selection that shape how organisms respond to their environments, as well as for the detection and characterisation of viral variants and the development of improved diagnostic tools. However, the analysis relied on alignment methods that compare sequences by identifying homologous regions, which poses a computational bottleneck. Thus, alignment-free techniques have the potential to maintain the quality of results whilst addressing scalability and explainability in sequence analysis.

Results: This study presents GREAC (Genomic Region Extraction and Classifier), a novel k-mer-based computational methodology for identifying discriminative genomic regions that reduces dimensionality while improving classification performance. The results indicate that selected k-mers are suitable information sources for viral genome analysis, including mutation behaviour, by comparing k-mer frequency profiles and classification methodologies. The GREAC method also exhibits robust performance on classification tasks, achieving high accuracy

and demonstrating strong data generalisation across varying training set sizes. Besides, the GREAC effectively reduces dimensionality by identifying discriminatory regions within biological sequences. In organisms such as SARS-CoV-2 and Monkeypox, dimensionality reduction is evident through the selection of specific coding genomic regions that retain the most informative mutations. These regions can reveal signals of speciation or population differentiation, helping to understand the evolutionary mechanisms behind divergence.
Conclusions: Focusing on informative regions rather than entire sequences, GREAC addresses the critical need for explainability and transparency in genomic results. Compared to competitor methods, GREAC offers transparency and explainability in its pattern discovery by directly extracting discriminative k-mers from raw genomic data and transforming them into representative signals. This strategy enhances the interpretability of genomic features while eliminating dependence on reference genomes. GREAC's open-source implementation ensures replicability and scalability, offering a robust tool for bioinformatics research and the development of tailored biotechnology applications, openly accessible at https://github.com/SALIPE/GREAC.

Keywords: Bioinformatics, pattern recognition, viral genomics, K-mer analysis, discriminative regions, alignment-free analysis, data-driven, dimensionality reduction

1 Background

The analysis of genomic sequences is a fundamental task in bioinformatics, providing critical insights into species' diversity [1]. With the continuous advancement of high-throughput sequencing technologies, an ever-increasing amount of genomic data is being generated. This data deluge opens unprecedented opportunities for research on genetic variants, disease associations, and the identification of genomic signatures of natural selection that shape how organisms respond to their environments, as well as for the detection and characterisation of viral variants and the development of improved diagnostic tools [2]. However, extracting meaningful knowledge from these vast and complex datasets presents significant computational challenges [3, 4].

In general, studies based on genomic data aim to extract meaningful information from large-scale, high-dimensional datasets by treating them as a whole [5]. Thus, a primary challenge is that genome-wide DNA sequences, transcriptome-wide RNA sequences or protein sequences can be thousands or even millions of nucleotides or amino acids long, resulting in a feature space with an extremely large number of dimensions during computational analysis. This phenomenon, often referred to as "the curse of dimensionality" [6], can seriously compromise the performance of machine learning algorithms by increasing computational complexity, demanding more storage, and elevating the risk of overfitting, and in extreme cases, leading the model to learn noise instead of the true pattern signal [7]. Consequently, dimensionality reduction has become an indispensable step in the bioinformatics approaches, aiming to lead high-dimensional data into a meaningful, low-dimensional representation without losing critical information [3, 8–12].

Traditionally, biological sequence analysis has relied on alignment-based methods, which compare sequences by identifying homologous regions [13]. While these methods are powerful, they typically use computationally intensive techniques such as dynamic programming, resulting in high computational costs that scale poorly with increasing sequence size and number, making their application to large-scale genomic datasets usually infeasible or often impractical. To overcome this limitation, various developments of alignment-free methods have been proposed [14, 15].

These approaches combine techniques from data mining, feature extraction, signal processing, and machine learning to analyse biological sequences, including converting them into numerical vectors (e.g., based on k-mer frequencies), which enables efficient comparison and classification. Alignment-free methods, and k-mers in particular, provide powerful and versatile solutions for multiple types of biological sequence analysis [16], including species classification [17], RNA identification [18–20], phylogenetic reconstruction [21, 22], and metagenomic analysis [23].

The use of k-mer frequencies as an informational feature has been adopted in previous studies to reveal composition spectra, such as non-random oligonucleotide usage driven by GC content and CpG suppression, and structural patterns, including the emergence of unimodal versus multimodal k-mer frequency distributions across taxa [17, 24]. For instance, CASTOR-KRFE [12] and KEVOLVE [25] are alignment-free machine learning approaches that identify minimal, discriminative k-mer sets for viral (sub)variant classification. CASTOR KRFE employs Recursive Feature Elimination (RFE) with a linear Support Vector Machine (SVM) to select a reduced feature subset that maximises classification performance. In contrast, KEVOLVE uses a genetic algorithm–driven search, coupled with a machine-learning kernel, to discover minimal k-mer ensembles. In addition, KANALYZER [26] enables the identification of k-mers in genomic regions, leading to signatures in genomic sequences (genomic locations, nucleotide variants, and amino acid effects).

Although alignment-free methods successfully bypass the computational bottleneck of traditional sequence alignment algorithms, their conversion of sequences into feature vectors often increases the problem of high dimensionality [1]. Therefore, there is an urgent need for methods that effectively reduce the dimensionality of these feature vectors without losing crucial biological information.

In this context, we introduce GREAC (Genomic Region Extraction and Classifier), a novel data-driven method for dimensionality reduction and classification of biological sequences. Our approach distinguishes itself not only by reducing the feature space but also by strategically selecting only the most informative regions of the sequences that contribute most to the classification task. By focusing the analysis on these key regions, our method aims to improve classification accuracy, improve computational efficiency, and provide more interpretable results.

The proposed method is designed around three distinct phases: 1) extraction of variant-exclusive k-mer sets derived from the input data; 2) identification of exclusive genomic regions containing these k-mers, achieved by interpreting the genome as a signal and filtering segments based on the presence of k-mer to reduce dimensionality; and 3) construction of a reference signal using k-mer occurrence frequencies within each

region to establish a signature for variant differentiation and classification. This comprehensive pattern recognition workflow renders the method inherently data-driven. Using a sliding-window approach with specific k-mers, the algorithm avoids *a priori* assumptions, modelling its behaviour directly from the observed data features [27].

The GREAC's suitability was validated by analysing its results across five viruses – SARS-CoV, DENV, HBV, HIV and Monkeypox – and comparing them with those of similar methods reported in the literature. This assessment employs a rigorous methodological approach, corresponding to the worst-case application scenario, where viruses exhibit the highest known mutation and recombination rates among organisms, posing a computational challenge in which detecting generalist patterns becomes particularly complex [28–30].

The application and validation of computational methodologies on viral datasets is justified, given the large volume of existing data for a high number of potential pathogens, which, according to the World Health Organisation (WHO), vastly exceeds the available resources for disease research and development. However, like other alignment-free methods, the GREAC method can be applied to a wide range of genomic analyses, including those involving eukaryotic species.

The GREAC method is implemented as open-source software, designed to be easily adaptable through community-driven modifications. This accessibility enables researchers to customise the approach according to their specific analytical requirements and research objectives. The software is freely available on GitHub [31], along with its source code, data, and usage examples, enabling replication of this study and facilitating adoption of the method.

2 Methods

2.1 Datasets

Evaluating methodologies requires careful selection of datasets, as data quality directly influences results: high-quality datasets yield reliable outcomes, while poor-quality data compromise analytical validity. To ensure reproducibility and enable meaningful comparison with other methodologies, we used publicly accessible datasets and an open-source approach, enabling replicable execution and providing a common foundation for benchmarking within the scientific community.

Table 1 summarises the viral datasets used in this study, which include five epidemiologically significant pathogens: SARS-CoV-2 (severe acute respiratory syndrome coronavirus 2), DENV (dengue virus), HBV (hepatitis B virus), Monkeypox (Orthopoxvirus monkeypox) and HIV-1 (human immunodeficiency virus type 1). This diverse collection enables a comprehensive evaluation of methodologies across viral genomes that differ in type, size, mutation rate, and evolutionary dynamics, providing a robust framework for assessing the generalizability and performance of our computational approaches across various viral families and genomic characteristics.

The selected datasets were obtained from previous studies and public repositories, allowing direct comparative evaluation across different methodologies. The SARS-CoV-2 dataset corresponds to the same collection used in KEVOLVE [25], the HIV-1 came from CASTOR-KRFE [12] and the DENV dataset was derived from

GRAMEP [11]. This approach ensures methodological consistency and allows for fair comparison of classification performance across established benchmarks in viral genomics research. The HBV sequences, representing six genotypes, were obtained from HBVdb [32], while the Monkeypox sequences comprising all complete genomes from both lineages were retrieved from the NCBI Virus database.

Table 1: Datasets Overview

Dataset (Source)	Variant	Average Length (σ)	Sequence Amount
HIV (CASTOR-KRFE)	HIV1_A	8,878 (438.02)	133
	HIV1_B	9,356 (439.44)	130
	HIV1_C	9,148 (349.10)	130
	HIV1_D	8,966 (381.55)	89
	HIV1_F	8,953 (361.35)	91
	HIV1_G	9,041 (363.94)	94
	Total		667
SARS-CoV-2 (KEVOLVE)	Alpha	29,838 (55.72)	175,212
	Beta	29,810 (56.96)	695
	Delta	29,801 (61.85)	9,408
	Epsilon	29,834 (63.59)	14,674
	Eta	29,741 (70.91)	716
	Gamma	29,810 (63.19)	8,129
	Iota	29,802 (56.11)	19,274
	Kappa	29,841 (52.04)	127
	Lambda	29,793 (55.19)	428
	Omicron	29,762 (65.99)	106,293
	Total		334,956
DENV (GRAMEP)	Type 1	10,585 (196.02)	2,571
	Type 2	10,647 (137.46)	1,756
	Type 3	10,647 (137.46)	1,272
	Type 4	10,422 (266.50)	462
	Total		6,061
HBV	A	3,220 (5.14)	818
	B	3,215 (2.48)	1,790
	C	3,213 (6.24)	2,724
	D	3,182 (3.32)	1,129
	E	3,211 (3.88)	298
	F	3,214 (3.84)	187
	Total		6,946
MONKEYPOX	I	196,058 (1537.40)	819
	II	196,952 (1482.09)	2,727
	Total		3,546

2.2 GREAC - Genomic Region Extraction and Classifier

The GREAC software is fully coded in the Julia Language [33], a multi-threaded, high-performance, dynamically typed and compiled programming language, optimised for scientific computing. GREAC is distributed as an open-source tool (source code publicly available) that runs through the Julia interactive command-line REPL (read-eval-print loop) built into the Julia executable, providing flexibility for various deployment scenarios. Figure 1 shows an overview of GREAC's stages and functions.

GREAC offers five main built-in functions, along with shell script example files in its repository: *extract-features, fit-parameters, benchmark, file-classification*, and *fasta-regions*. The application operates on structured data directories following the format specified in the documentation. During execution, GREAC generates model files and optimised parameters, which are stored in cached files for efficient reuse.

The ***extract-features*** function implements the main methodology for extracting regions and generating reference signals for subsequent classification and genome behaviour analysis. The ***fit-parameters*** function performs a comprehensive brute-force parameter optimisation to identify optimal classification parameters, including frequency threshold values for region consideration and window percentage sizes.

The ***benchmark*** function trains the model and creates a classification data structure. In more detail, the discriminative regions are extracted, and reference signals are created (as in the *extract-features* function). In addition, the performance is evaluated on test sequences, and sequences are classified, with F1-score, precision, and recall measured. As a result, the function outputs a confusion matrix in a `.csv` file, and another `.csv` with per-class probability that leads to the final classification for each sequence, along with a `.pdf` file containing the execution summary with the classification metrics and the graphs with the respective frequency signals.

The ***file-classification*** function performs a single classification for each sequence in the input sequences file, using the pre-trained model. The ***fasta-regions*** function returns a `.bed` file with three columns: organism name, start position and end position. Furthermore, if the dataset input file was specified, the reduction is applied for each variant class file, selecting only the extracted regions' positions in each sequence and outputting a `*_extracted.fasta` file.

The GREAC repository includes extensive documentation covering proper usage procedures, Jupyter notebooks that demonstrate parameter analysis during feature extraction and modelling, shell scripts that provide practical examples of executing source code, and a toy model example based on the same HBV dataset used in this study. This comprehensive documentation suite ensures accessibility for users of different experience levels and facilitates reproducible research workflows.

2.2.1 Region extraction based on k-mer appearance

Identifying discriminative regions within genomic sequences requires establishing variant-specific information sources that can distinguish between different variants. K-mer analysis represents a well-established genomic discrimination approach that employs word-based methodologies, similar to single-nucleotide polymorphisms

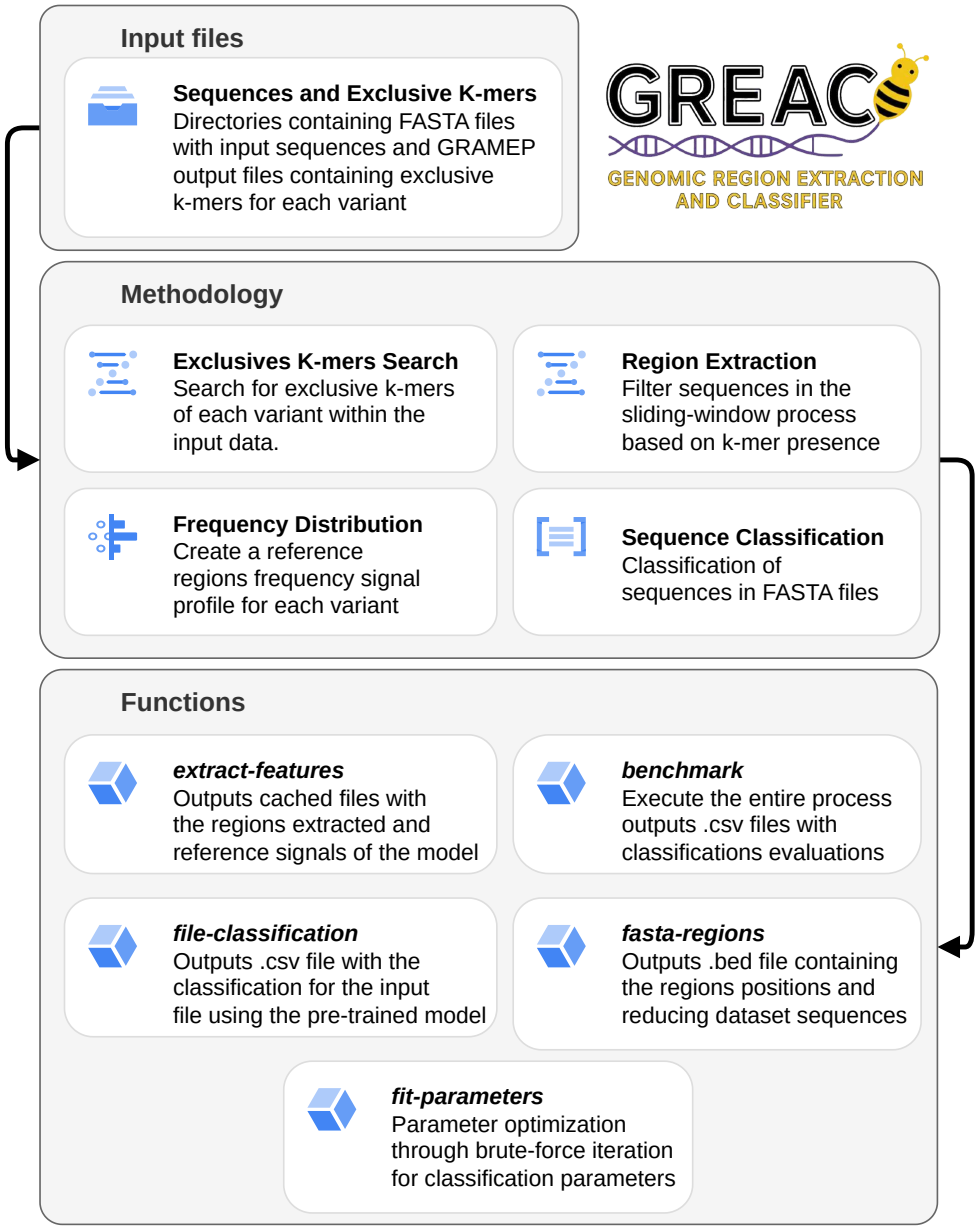


Fig. 1: Overview of GREAC.

(SNPs) but with a substantially larger information set, thereby providing enhanced discriminative power for sequence classification [34, 35].

The identification of discriminative k-mers is performed solely from the inherent information in the input datasets, eliminating the need for a reference sequence and

thereby removing any bias introduced during k-mer extraction that might affect the results. The method seeks only k-mers shared among all sequences within a given variant set and selects the most informative ones through a maximum entropy function, similar to the approach used in GRAMEP [11], an alignment-free open-source tool that applies Shannon entropy principles to identify regions of variation among samples compared to a reference genome and extract all variant-exclusive k-mers from them.

After identifying the exclusive k-mer sets, all intersections are removed, retaining only those k-mers uniquely exclusive to each variant in the sequence dataset. The entire selection process is illustrated in Figure 2.

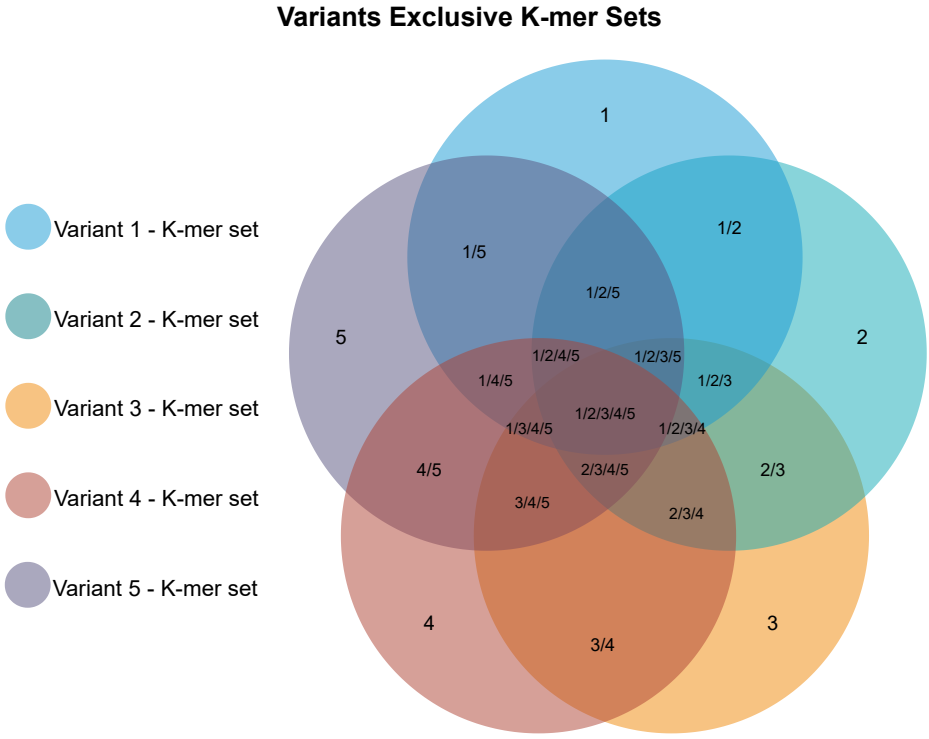


Fig. 2: Venn diagram illustrating the exclusive k-mer selection process among five distinct variants. Exclusive k-mers are extracted from each sequence set, and only those that are uniquely exclusive to individual variants (represented by 1, 2, 4, and 5) are selected, excluding any k-mers shared between variant sets at any level.

Region identification was performed individually across variant datasets using an accurate genome-wide search, simulating a sliding-window approach to enhance performance. The sliding-window methodology represents a well-established approach in related previous studies and serves as a standard algorithm for data-driven genomic

analyses [9–11, 36]. The window size calculation is formalised in Equation 1, where $|\mathbb{S}|$ is the vector containing all lengths of the sequences, and P is the selected percentage value, thus determining the window size based on the shortest sequence length.

$$W(|\mathbb{S}|, P) = argmin(|\mathbb{S}|) \times P. \tag{1}$$

During the sliding-window process, each window with at least one of the k-mers from the variant-specific set increments a positional counter. After processing all training samples and traversing the entire genome, the counter generates a frequency histogram, with peaks indicating regions with high concentrations of discriminative k-mers. Several of these histograms, for different viruses, are shown in Figure 3.

The selected regions from the histograms have a threshold parameter that can be set to benchmark performance or to specify a higher minimum appearance percentage across all training samples to be considered. With the selected windows, it is possible to retrieve all nucleotide positions and extract variant-specific regions.

To mitigate the impact of sequence length deviation, we incorporate a safety margin during region extraction. This boundary is adjusted based on the variance of the sequences and the window dimensions, providing an error-tolerant buffer that preserves discriminative information despite minor length inconsistencies, while accounting for positions that could be lost due to these variations.

It is necessary to combine all variant-specific regions into a unified positional conjunction as shown in Figure 4. This process enables a fair comparison across variants, ensuring consistent regional representation while preserving the discriminative characteristics identified through k-mer frequency analysis.

2.2.2 Using k-mers as information over EIIP values

The selection of k-mers as the primary source of information for region extraction and subsequent pattern discovery was not arbitrary; it was grounded in related research. Previous studies have demonstrated the effectiveness of numerical representations for genomic sequences, particularly through the use of Electron-Ion Interaction Potential (EIIP) values, where each amino acid is assigned a specific numerical value [37, 38]. This approach extends to the translation from DNA to RNA and, subsequently, to the primary protein sequence, with each receiving its corresponding EIIP value.

The use of numerical series to represent amino acid sequences enables the application of digital signal processing methods, such as Resonant Recognition Model (RRM) [37]. This methodology involves converting primary protein sequences into EIIP numerical series and then transforming them into the frequency domain using the Fast Fourier Transform (FFT). The multiplication of these signals (*cross-spectrum* function) serves to extract the most discriminative information regarding their biological functionalities.

Previous research employing this methodology has yielded promising results when applied to genomic data from diverse organisms. Specifically, studies that used RNA reads from the PLEK [39] and CPC2 [40] tool datasets have successfully classified non-coding RNA (ncRNA) and messenger RNA (mRNA) sequences [20].

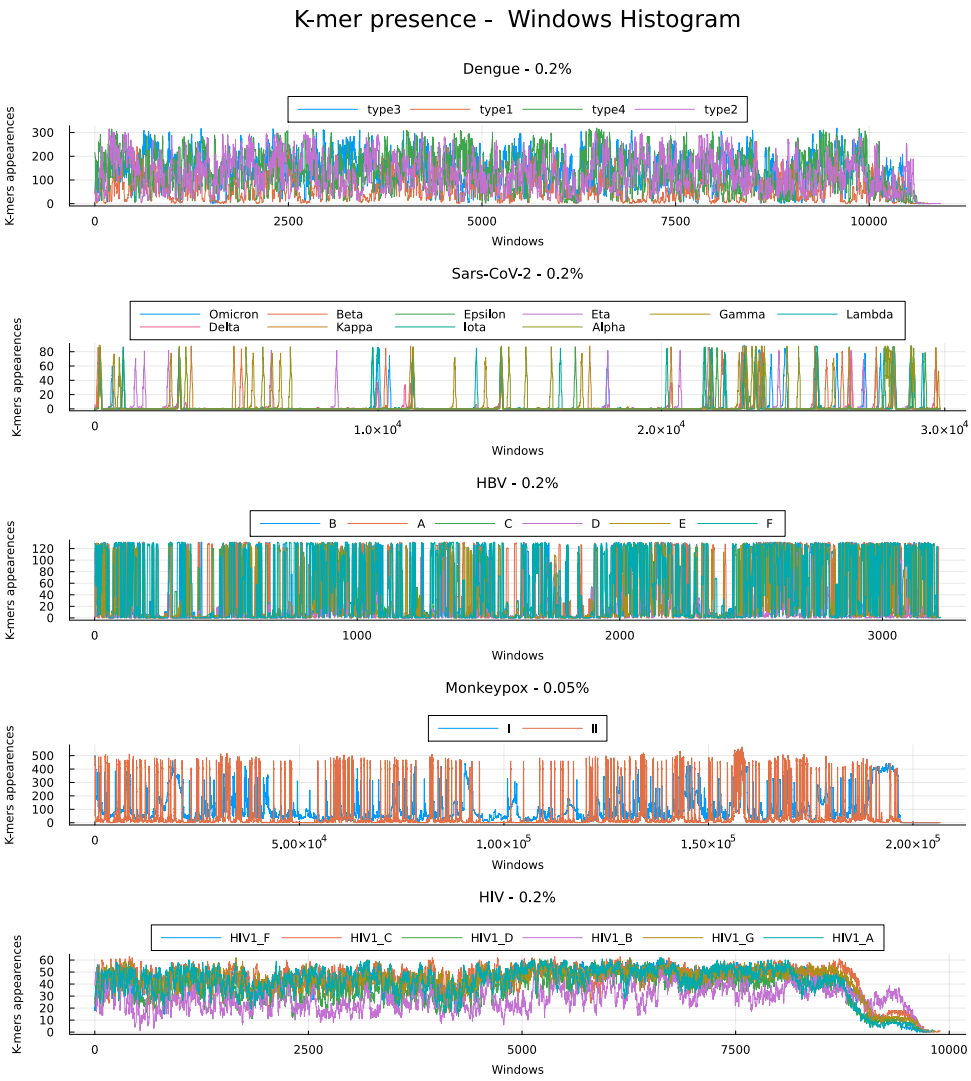


Fig. 3: Histograms of the presence of an exclusive k-mer in each region among all the training samples. Each chart corresponds to a different virus histogram, where the x-axis represents the windows and the y-axis the count of how many of the training samples have the presence of one of the k-mers in that region.

The transition from EIIP to k-mer representation was motivated by two fundamental limitations of the RRM approach when applied to direct sequence position identification. First, the frequencies extracted through RRM that represent biological functions do not correspond to direct positions in DNA sequences, as these frequencies only indicate the possibility of applying filters that would transform protein

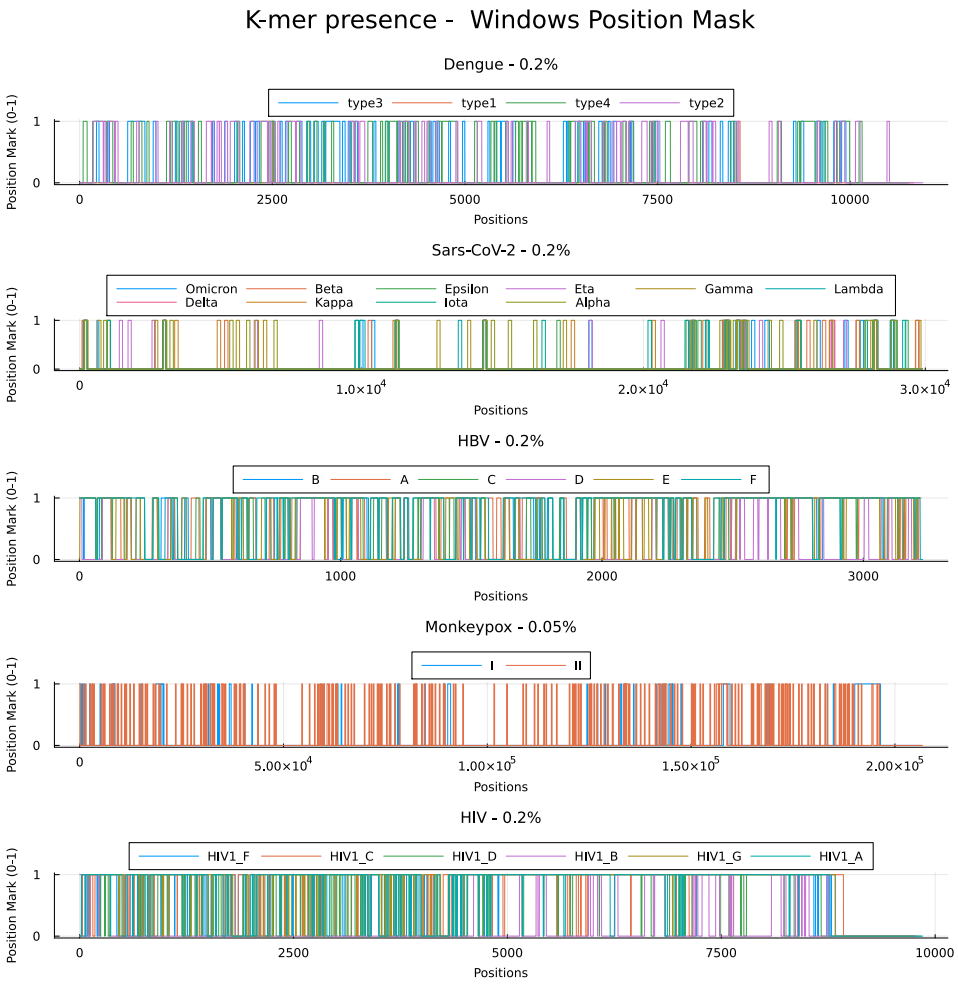


Fig. 4: Extracted regions' positions in each variant by histogram. Each chart shows, respectively, the position mask for each virus, where the x-axis represents the sequence position and the y-axis represents whether the position is part of the region.

sequence signals into sinusoidal waves. Second, and more critically, there is no feasible method, to the best of our knowledge, for reverse-translating protein sequences back to nucleotide sequences, since the genetic code represents a non-bijective function where each codon triplet codes for a specific amino acid, but multiple codons can code for the same amino acid. Any solution would require alignment with the original sequence, which contradicts the principle of alignment-free methodology and would introduce computational complexity far beyond the scope of this study. Therefore, k-mers provide a more direct and tractable approach to positional analysis while maintaining the numerical processing advantages of the established methodologies.

2.2.3 K-mers frequency distribution reference

With the region's positions extracted, we aggregated frequency values based on the *relative frequency*, which is used to calculate the number of times the event occurs in a given set of possible occurrences [41]. Those sets comprise all variants, exclusive of k-mers, the same used for the region extraction stage, treating all variants as possible discriminative regions within variant sequences, focusing on better generalisation.

The relative frequency of the regions is measured using Equation 2, considering all variant training sequences, where K is the k-mer set, $|K|$ denotes the total k-mers, and N is the number of samples. The numerator is the sum of $count(K)$ across the samples, which counts the number of k-mers that appear in each region. This way produces a normalised frequency for each wr region, as follows:

$$F_{wr} = \frac{1}{N \times |K|} \sum_{i=0}^{N} count(K)_i \qquad wr = 1, 2, 3, ..., R. \tag{2}$$

Using the relative frequencies of the regions for each variant, a reference signal profile is constructed, representing variant-specific frequency patterns and showing the organism's behaviour, as demonstrated in Figure 5.

2.2.4 Classification

With established reference values characterising the behaviour of the variants within the analysed regions, we return to the training data to calculate the mean and standard deviation of the observed distances. This process allows for the definition of confidence intervals, conceptually similar to those utilised in cluster-based analysis. Furthermore, the specific k-mers comprising these calculated signals are recorded; consequently, the methodology preserves both the numerical signals and their underlying metadata, which together enhance the robustness of the classification process.

To quantify sequence similarity within extracted regions, we evaluated multiple distance metrics, including Euclidean, Manhattan, and Chi-Squared distances [42]. Following empirical testing across all viral datasets, the Manhattan distance (Equation 3) was selected for its superior performance and computational versatility across diverse genomic datasets.

$$D(Q, P) = \sum_{i=1}^{n} |Q_i - P_i| \tag{3}$$

In multiple-class problems, the fuzzy sets approach uses membership functions to assign degrees of membership based on distance values. Membership degrees indicate how closely the new input values correspond to each class [43]. The Gaussian Membership function $\mu(d)$, calculated using Equation 4, is used to determine the sequence referenced in [44], where d represents the calculated distance between the input sequence and the variant reference distribution, and μ and σ are the mean and standard deviation distances within the variant group.

$$\mu(d) = \exp\left(-\frac{(d-\mu)^2}{2\sigma^2}\right) \tag{4}$$

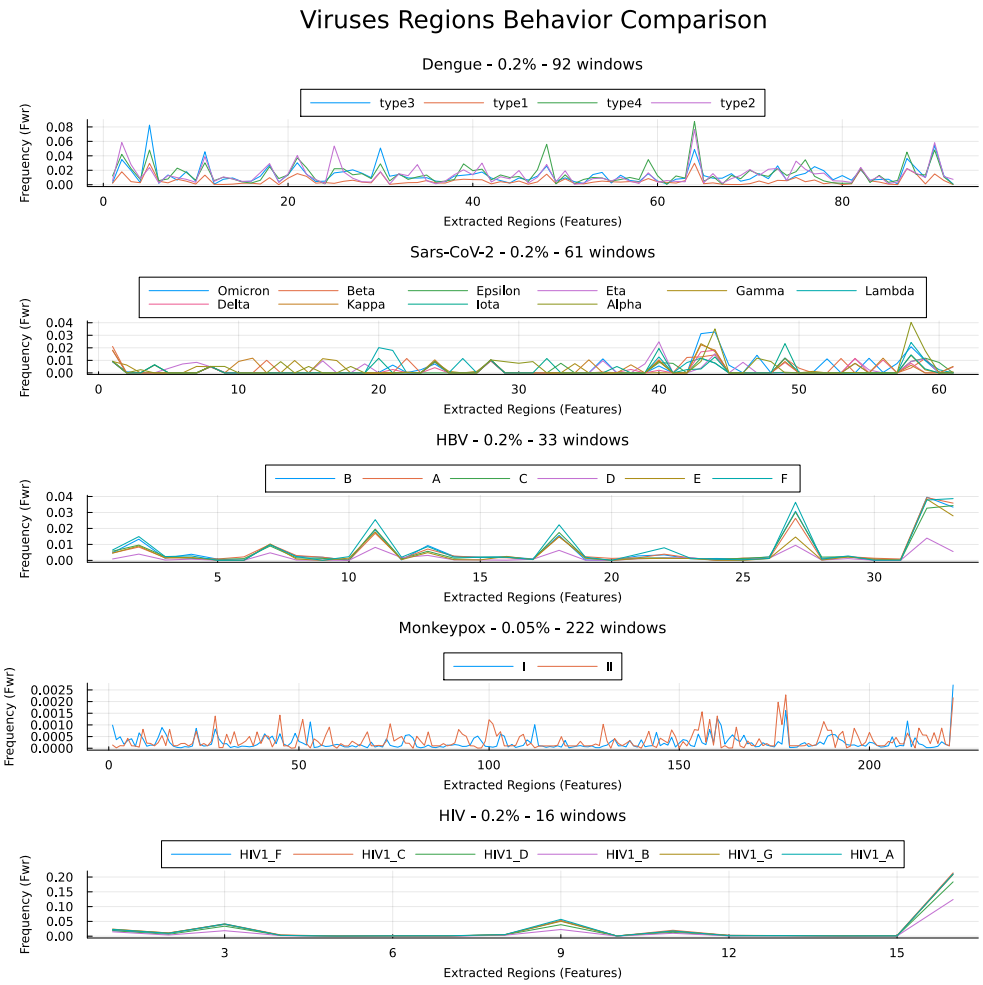


Fig. 5: Reference behaviour signals created using Equation 2. Each chart shows the behaviour of each virus variant, with the x-axis representing the extracted regions and the y-axis representing the relative frequency of k-mers.

The formula outputs values between $[0, 1]$ for each variant reference signal, which makes it possible for classification applying the *argmax* method to select the variant with the highest degree of membership or to be used as a feature in classifiers. In experiments conducted during the development phase, we concluded that these membership values exhibit discriminative characteristics that enhance the classifier's performance.

Similarly to Gaussian membership, other features can also be extracted from the reference signals generated by the method in order to incorporate training features, such as the measured frequency signals from the sequences, the previously calculated distance, membership values between comparative reference signals and a *signal feature* that captures variations between windows. The last one is a feature that is a

vector calculated using Equation 5, where $diverg[n]$ represents the divergence signal between the relative frequency of the k-mer set of the current window and the consecutive window of the signal x.

The aggregation of this signal in the model's feature vector enhances the classificatory power, allowing for the understanding of internal variation between signal windows as a behavioural pattern across frequency values.

$$diverg[n] = x[n+1] - x[n] \tag{5}$$

Random Forest, a tree-based classifier, was chosen for the final classification assignment, implemented to improve classification accuracy through more complex, high-variance models, enabling a more comprehensive analysis of the components that make up the signals while maintaining the methodology's performance and explicability. For model training, we used the default implementation parameters, setting a multi-class classification with the *softprob* activation function. This configuration produces output probabilities for each class, allowing a more granular analysis of the uncertainty associated with each prediction.

The classification process was performed for each variant available in the dataset. One of the main characteristics is the distance between the F_{wr} signal corresponding to each class, the Gaussian membership, and the window frequency signal of the input data to be classified. For each individual classification, we extract the probability assigned by the classifier. Subsequently, we apply the $argmax$ function on all probabilities obtained to determine the final class with the highest confidence. This process involves assembling the set of probabilities from *softprob* classifications into a probability matrix $\mathbf{P}_{n\times n}$, where n is the number of classes.

$$\mathbf{P} = \begin{pmatrix} P_{11} & P_{12} & \cdots & P_{1n} \\ P_{21} & P_{22} & \cdots & P_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ P_{n1} & P_{n2} & \cdots & P_{nn} \end{pmatrix} \quad \mathbf{d} = \begin{pmatrix} P_{11} \\ P_{22} \\ \vdots \\ P_{nn} \end{pmatrix} \tag{6}$$

In matrix $\mathbf{P}$ (Equation 6), each row i represents the probabilities when the model iterates over class j, and P_{ij} is the probability that the sample belongs to class j when iterating over class i (diagonal element), where $\sum_{j=1}^{n} P_{ij} = 1$ for each row i. Thus, we can extract the diagonal vector $\mathbf{d}$ and finalise the classification through $argmax$ of the diagonal vector where $\hat{y}$ is the predicted class, as shown in Equation 7.

$$\hat{y} = \underset{i\in\{1,2,\ldots,n\}}{argmax} \; P_{ii} \tag{7}$$

2.2.5 Evaluation Metrics

For an accurate evaluation of results, a confusion matrix provides a comprehensive view of the classifier's behaviour by showing the relationship between actual viral types (true labels) and predicted classifications. For better visualisation, the values are normalised, where each cell represents the percentage of samples from a true viral

type (rows) that were classified as a particular predicted type (columns). The diagonal elements indicate correct classification rates, while the off-diagonal elements reveal specific misclassification patterns between viral genome types.

The results represent means and standard deviations of 100 independent executions, demonstrating statistical reliability across multiple validation runs and accounting for stochastic variations in the classification process. The training data were randomly shuffled for each iteration, and the complete data processing pipeline was executed independently in each run, eliminating potential bias from data ordering and ensuring that the observed performance metrics reflect the true generalisation capability of the methodology.

Other metrics, such as *precision*, *recall*, and *F1-Score*, defined in Equations 8–10, respectively, are complementary. For multiclass classification, they were calculated as unweighted averages across all classes to mitigate bias towards prevalent variants.

$$Precision = \frac{1}{N}\sum_{i=0}^{N}\frac{TP_i}{TP_i + FP_i} \tag{8}$$

$$Recall = \frac{1}{N}\sum_{i=0}^{N}\frac{TP_i}{TP_i + FN_i} \tag{9}$$

$$F1 - score = \frac{1}{N}\sum_{i=0}^{N} 2 \times \frac{Precision_i \times Recall_i}{Precision_i + Recall_i} \tag{10}$$

3 Results

3.1 Classification Performance and Dimensionality Reduction

The computational experiments and their corresponding results are described in the following sections, beginning with data pre-analysis and data understanding. Because GREAC is a data-driven method, an initial analysis step is necessary to select the optimal parameters for each dataset. Given biological differences between organisms, such as sequence lengths and variation rate, adjustments are necessary in k-mer size (k), window percentage, and appearance threshold to achieve optimal performance. Finally, we demonstrate the results in both data dimensionality reduction and classification accuracy across all viral genome datasets.

The feature extraction methodology identifies discriminative regions between variants by searching for exclusive k-mers for each variant in the input dataset. The search and selection process (described in detail in Section 2.2.1) is performed during the "model training" process. This approach is conceptually similar to that used in GRAMEP [11], employing the maximisation of Shannon entropy to select the most informative k-mers and generate a set of k-mers from training samples for subsequent use in region extraction, without a reference sequence.

The discriminative regions are identified using a sliding window (whose size is a percentage of the minimal sequence length of the variant sequences) and searching for

the presence of any exclusive k-mers within the traversed windows, operating similarly to a filter. Applying this filter to all sequences yields a histogram of the most informative regions in the dataset, with selection based on a threshold parameter.

Defining the optimal value of k in alignment-free analysis remains a significant challenge, particularly due to divergent sequence lengths. Typically, in k-mer extraction tools such as GRAMEP, higher k values generate larger information sets. However, this increase in complexity can hinder the objective of identifying specific genomic regions. In our approach, higher k values yield smaller k-mer sets, i.e., longer sequences tend to be more specific, as we select exclusively from the input dataset without aligning or comparing to reference sequences to avoid bias.

Therefore, determining the minimum k value that generates exclusive information sets is fundamental. While similar alignment-free analysis methods have employed reduced-size k-mers in variant analysis [45, 46], the optimal range is highly dependent on the analytical context. This variability is supported by previous research regarding the search for unique information sets [47].

Consequently, we evaluated two distinct odd-numbered k-mer sizes ($k = 5$ and $k = 7$) to mitigate issues associated with palindromic sequences, which are often encountered with even-numbered k-mers. As we target exclusive k-mers, increasing the size of k reduces both the quantity of exclusive k-mers and their probability of occurrence. This is particularly relevant for organisms with smaller genomes, such as DENV, HBV, and HIV, where the probability of finding unique sets is naturally lower. Our results indicated that $k = 5$ yielded excessively large datasets due to the higher probability of non-exclusive recurrence. In contrast, $k = 7$ not only presented the best results, but was also the most consistent value across tests, as demonstrated in Table 2, which shows the optimal results obtained from the parameter fitting tests using this specific k-mer size.

Table 2: Results of the computational experiments for the k-mer size evaluation (F1-Score).

k-mer Size	SARS-CoV-2	HIV	HBV	DENV	Monkeypox
5	91.82%	97.53%	99.98%	99.28%	97.49%
7	98.14%	99.56%	99.99%	99.73%	97.80%

For the remaining parameter selection, a brute-force approach was employed, systematically testing window-size percentages across different ranges, based on sequence length. The goal was to identify window sizes that maintain proportional relationships with k-mer sizes, with longer genome sequences using smaller percentages and shorter sequences using larger percentages, to achieve optimal feature extraction performance.

These parameter ranges were determined based on computational analyses, in which a window size range of 0.1% to 0.25% was employed for SARS-CoV-2 and DENV (dengue virus) sequences. The range is 0.15% to 0.2%, with increments of 0.05%. For shorter sequences, such as those of HIV and HBV (hepatitis B), the range starts at 0.15% and 0.2%, respectively, with increments of 0.05%. For Monkeypox,

which comprises significantly longer sequences, a more refined range, starting at 0.05%, was adopted. Simultaneously, the thresholds for the appearance of k-mer in windows among the sequences were evaluated in the range of 50% to 80%, with incremental steps of 5%, as detailed in the complementary data in Appendix A.

This process resembles Recursive Feature Elimination (RFE)[12]. However, it focuses on identifying optimal parameter values rather than on feature selection for classification models. To evaluate the parametrisation, Table 3 presents the results of the lead evaluation, assessing how each input parameter performed across organisms.

Table 3: Selected parameters from the performance evaluation

Dataset	**k-mer Size**	**Window Size (%)**	**Threshold (%)**
SARS-CoV-2	7	0.15	80
Monkeypox	7	0.15	60
HBV	7	0.20	60
HIV	7	0.15	70
DENV	7	0.15	75

Given the data imbalance of the datasets, it is necessary to consider a balancing approach to achieve a fair evaluation. Thus, a $70\% - 30\%$ train-test split was performed based on the variant with the fewest sequences, thereby equalising the training sample for all classes and testing with complements. The sequences composing the train and test sets were randomly selected to avoid bias. In this way, the method's generalisation capacity can be evaluated as a complement. Given the low number of samples for the classes with the fewest sequences, the training step was performed using a low-volume dataset: 63 samples for HIV, 89 samples for SARS-CoV-2, 131 for HBV, 324 for DENV and 574 for Monkeypox.

To accurately evaluate the classifier's performance, we primarily used confusion matrices to provide a comprehensive overview of classification results and the relationship between true and predicted classifications. To ensure statistical reliability and account for stochastic variations, the reported results represent the mean and standard deviation (σ) derived from 100 independent executions. In each iteration, the training data were randomly shuffled, and the complete processing pipeline was executed independently, thereby mitigating potential ordering bias and ensuring the metrics reflect the true generalisation capability of the methodology.

To validate the proposed approach against established methods from the literature, we adopt uniform comparative experiments using two genomic classification tools: CASTOR-KRFE and KEVOLVE. All viral datasets were subjected to identical training and testing scenarios to enable a direct comparison. The proposed method (GREAC) demonstrated robust performance across all analysed organisms. Table 4 presents the mean F1-score achieved along with the standard deviation for each dataset compared to other methodologies. The complementary confusion matrices with the results for each comparison evaluation are available in the supplementary files.

Table 4: Methods Mean F1-Score (σ) Comparison

Dataset	GREAC	KEVOLVE	CASTOR-KRFE
HIV	98.90% (±0.64%)	18.24% (±15.35%)	97.12% (±1.43%)
SARS-CoV-2	95.95% (±2.19%)	4.82% (±5.66%)	85.96% (±4.02%)
DENV	98.89% (±0.39%)	25.00% (±18.91%)	98.56% (±0.79%)
HBV	98.90% (±0.64%)	21.58% (±14.30%)	96.29% (±1.84%)
Monkeypox	96.00% (±1.10%)	82.87% (±34.53%)	99.68% (±0.33%)

3.2 Dimensionality Reduction

The GREAC region extraction methodology reduces dimensionality by prioritising discriminative genomic regions over full-sequence analysis. This targeted approach results in considerably less input data, thereby reducing processing and storage requirements and enabling greater scalability for sequence analysis and classification.

As evidenced in Table 5, the impact of this reduction is greater for larger viral genomes. Specifically, the method reduced the SARS-CoV-2 dataset by approximately two-thirds (≈67%) and the Monkeypox dataset by 72%. These figures translate into practical storage benefits, with uncompressed raw FASTA files showing size reductions of approximately 68% for SARS-CoV-2 and 74% for Monkeypox. In contrast, the reduction observed in smaller genomes was marginal, amounting to approximately 100 nucleotides. This variation highlights the method's ability to selectively compress data where redundancy exists while preserving the dense information content typical of compact viral genomes.

Table 5: Dimensionality reduction results across datasets

Organism	Full Genome (Median)	Reduction Achieved (Median (σ))	(%)
SARS-CoV-2	≈29803 bp	9587.85 (53.64)	67.83
Monkeypox	≈196505 bp	54588.00 (0.00)	72.22
DENV	≈10575 bp	10491.30 (382.67)	0.79
HBV	≈3209 bp	3120.51 (17.14)	2.77
HIV	≈9057 bp	8857.62 (58.16)	2.20

3.3 Alternative k-mer Sources and Classification Strategies

K-mer sets derived from GRAMEP [11] were used in the early development of GREAC, showcasing the methodology's versatility with third-party input data. However, the final version yielded superior results and performance using k-mers extracted solely from the input data, without relying on external reference sequences, thus following a strictly data-driven approach. Despite this preference for internal extraction, full support for GRAMEP-derived k-mers has been maintained and is natively implemented as a software option, allowing users to leverage external k-mer sets when necessary.

3.4 Classification Strategies

To validate GREAC and its suitability in classifying viral genomes, two approaches were considered. The first employs Gaussian membership function values to assign probabilistic belonging scores, providing a soft classification framework that captures uncertainty in boundary cases. The second approach employs a tree-based classifier to assess discrimination and classification while maintaining the explicability that the method inherently provides. This dual approach enables robust performance across various classification scenarios while maintaining the interpretability of the results.

3.4.1 Gaussian Membership and Classification

The first experiment was performed to evaluate the classification performance of fuzzy Gaussian membership classification, thereby validating the potential of the reference behaviour signal generated by the GREAC method. Table 6 presents the mean F1-score achieved and the standard deviation for the adopted datasets.

Table 6: GREAC Gaussian Results - Mean F1-Score (σ)

HIV	**SARS-CoV-2**	**DENV**	**HBV**	**Monkeypox**
51.98% (±3.05%)	94.32% (±2.81%)	98.20% (±0.86%)	99.87% (±0.09%)	77.47% (±0.64%)

The HIV classification was the most challenging due to the high variability in sequence length (approximately 4.3% of the genome), which significantly affects the position-based methodology. The substantial standard deviation in sequence length disrupts the positional consistency required for accurate variant identification, resulting in lower classification performance than in other viral organisms.

The classification results obtained with SARS-CoV-2 demonstrate the method's versatility when applied to viruses with varying genome sizes. Having low variation in sequence lengths (representing $\approx 0.2\%$ of the genome), such as DENV (representing approximately 1.7% of the genome), results in consistent classification with high accuracy values.

The HBV dataset classification achieved consistent accuracies across all variants. Benefiting from its compact genome and low sequence-length variability (approximately 0.1% of the genome), the classification process yielded accurate and consistent regional identification and discrimination when applied to new input sequences.

The Monkeypox dataset comprises the longest sequences among all datasets tested, highlighting the methodology's versatility and scalability for feature extraction across sequences of varying sizes. This characteristic offers valuable insights into the method's adaptability across diverse genomic contexts and sequence complexities.

3.4.2 XGBoost Classification

To further improve multi-class classification performance, features derived from genomic regions extracted by GREAC were used to construct input vectors for an

XGBoost classifier [48]. This approach augments the methodology with enhanced predictive capabilities while preserving explainability and transparency, as XGBoost is fundamentally based on decision tree ensembles. To enrich the feature vectors and optimise the results, the Jaccard similarity between each extracted region and the reference sequence is calculated using MinHash. This probabilistic hashing algorithm is specifically designed to rapidly estimate similarity across high-dimensional datasets by reducing them to compact sketch representations, thereby efficiently capturing relationships of approximate equality and containment between sets [49, 50].

The classification experiments used the same parameters established in the previous experiments, with the final prediction assigned to the class with the highest probability. As shown in Table 7, integrating these features significantly improved overall classification performance, achieving high precision across the test datasets.

These experiments demonstrated the efficacy of employing tree-based classifiers to improve predictive accuracy. However, despite XGBoost's strong initial performance, the final software implementation switched to a Random Forest algorithm due to deployment and distribution constraints. Because XGBoost is an external dependency, it introduced compilation errors during software distribution. By replacing it with a Random Forest classifier, which is natively supported in our development environment, we eliminated these compilation bottlenecks and improved overall execution performance while maintaining comparable and highly consistent classification results.

Table 7: GREAC XGBoost Results - Mean F1-Score (σ)

HIV	**SARS-CoV-2**	**DENV**	**HBV**	**Monkeypox**
83.32% (±2.51%)	92.74% (±3.35%)	98.03% (±0.61%)	98.48% (±0.85%)	95.67% (±1.55%)

3.4.3 Using reduced datasets with other methodologies

To provide additional validation and demonstrate that the dimensionality reduction yields regions with discriminatory power for classification, we applied the reduction directly to the data, using the extracted regions to construct a dataset in which sequences are reduced to comprise only these regions. The reduction was applied to the DENV, SARS-CoV-2, and Monkeypox datasets, which showed the most significant dimensionality reduction among datasets with medium sequence length.

Results from CASTOR-KRFE confirmed the approach's efficiency, yielding classification performance similar to that of the original full datasets. This stability is clearly shown in Table 8 for SARS-CoV-2, Monkeypox and DENV, proving that the discriminatory power is preserved even after substantial data reduction.

Furthermore, KEVOLVE tests confirmed that the reduction process does not result in a loss of precision, as overall results showed insignificant deviations from the baselines. However, a distinct improvement was observed in the Monkeypox dataset, where dimensionality reduction corrected inconsistencies in the original analysis, resulting in significantly higher accuracy and consistency than in the full dataset.

Table 8: Comparison of the methods Mean F1-Score (σ) with reduced dataset

Dataset	KEVOLVE	CASTOR-KRFE
SARS-CoV-2	18.19% (±20.11%)	85.62% (±3.88%)
DENV	23.86% (±16.16%)	98.35% (±0.00%)
Monkeypox	96.63% (±1.25%)	97.67% (±1.42%)

4 Discussion

The results of computational experiments indicate that k-mers are suitable information sources for the development of viral genome analysis, including mutation behaviour, by comparing k-mer frequency profiles and classification methodologies (using a tree-based classifier and fuzzy Gaussian membership) with potential for data reduction. The GREAC method results also exhibit robust performance when applied to classification machine-learning methods, such as the XGBoost classifier.

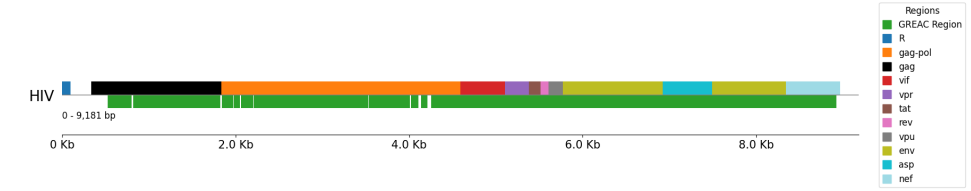
The comparative analysis reveals distinct performance patterns among the evaluated methods, with both GREAC and CASTOR-KRFE achieving high accuracy scores, demonstrating robust data generalisation across varying training set sizes.

In contrast, KEVOLVE exhibited significant performance limitations, characterised by low accuracy results and high standard deviation across experiments. These results suggest a limitation of the method: it requires larger volumes of training data to achieve comparable classification performance. These performance disparities can be attributed to the usage of a genetic algorithm that is susceptible to overfitting, especially when working with limited training data. The natural selection-based optimisation process, when constrained by insufficient samples, leads to inadequate feature selection and model over-specialisation to training data [51–53].

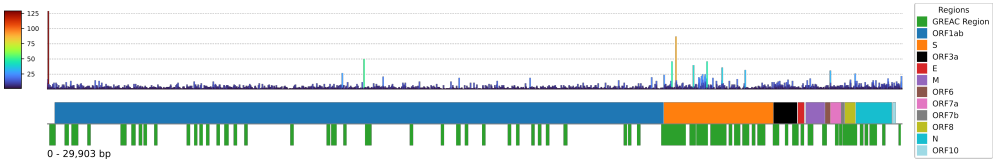
The GREAC method effectively reduces dimensionality by identifying discriminatory regions within biological sequences. An example of the regions obtained for the four organisms analysed is presented in Figure 6. In organisms such as SARS-CoV-2 and Monkeypox, the dimensionality reduction is evident through the selection of specific genomic regions. When analysing the origin of these extracted regions, it becomes clear that they predominantly belong to the coding sequences of the structural proteins of the SARS-CoV-2 virus, as exemplified by the method [54], thereby justifying its interpretative power.

Observing HBV, which has a circular genome, the identified regions are distributed throughout the genome, with a high incidence in the genic regions [55]. Similarly, in the Dengue virus genome, a single-stranded RNA genome, important windows are also found throughout the genome, a characteristic that can be attributed to the presence of crucial RNA secondary structures and a single open reading frame that encodes a large polyprotein, subsequently cleaved into individual functional proteins [56, 57].

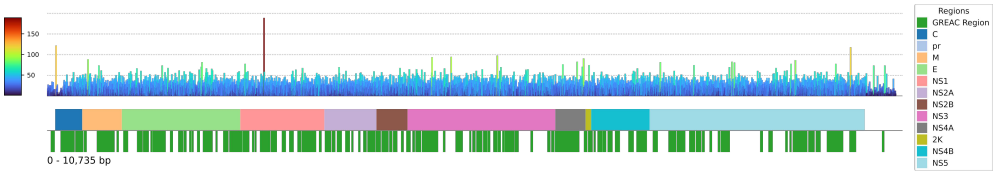
When comparing the regions extracted by GREAC with the high-frequency mutational events reported by Nextstrain [58], there is a clear overlap. This demonstrates the effectiveness of dimensionality reduction in retaining the most informative genomic regions from a mutational perspective. In addition to simplifying the data, this approach helps pinpoint highly mutable regions, which may be linked to the evolution



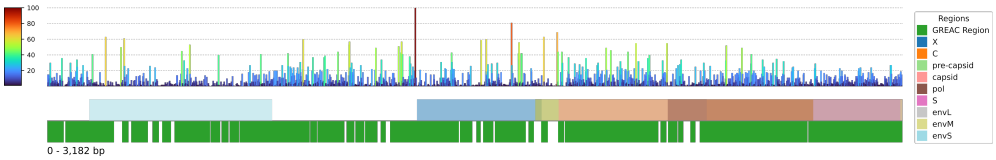
(a) HIV



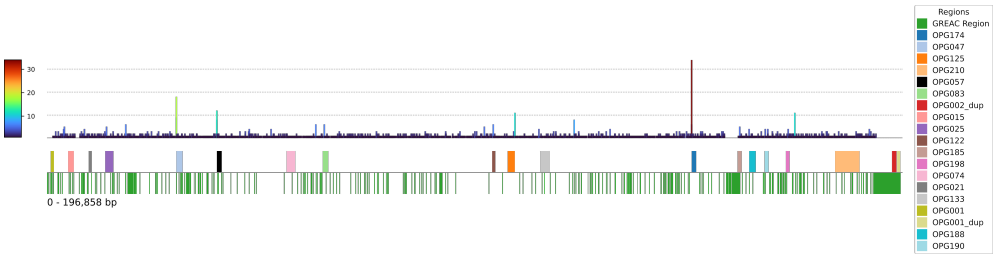
(b) SARS-CoV-2



(c) DENV



(d) HBV



(e) Monkeypox

Fig. 6: Genomic annotations and regions identified by the GREAC as important for different viruses: 6a) HIV (Human Immunodeficiency Virus), 6b) SARS-CoV-2, 6c) DENV (Dengue virus), 6d) HBV (Hepatitis B virus), and 6e) MPOX (Monkeypox). Known genomic annotations are shown in the upper tracks of each sub-figure. Regions highlighted as "important" are shown in the lower tracks, indicating overlap with previously annotated functional elements and revealing potentially relevant regions.

and adaptability of organisms [59]. These regions can reveal signals of speciation or population differentiation, helping to understand the evolutionary mechanisms behind divergence. Further biological interpretation is possible by investigating whether such regions coincide with regulatory elements (e.g. enhancers, promoters), transcription factor binding motifs, or structural features such as splice sites.

Compared to other analysed methods, GREAC distinguishes itself by offering greater transparency and explainability in its pattern-discovery methodology. By extracting discriminative k-mers directly from raw genomic data and transforming them into representative signals through mathematical modelling and Digital Signal Processing (DSP), the methodology introduces a novel layer of pattern recognition. This relationship between k-mer analysis and DSP not only enhances the interpretability of genomic features but also eliminates dependence on reference genomes, providing a scalable, transparent solution for biotechnology research.

5 Conclusions

The application of pattern recognition and machine learning methods for identifying, characterising, and classifying genomic mutations and variants is fundamental across modern biological sciences. These approaches enable deeper insights not only in healthcare and biotechnology, but also in evolutionary biology, ecology, conservation genomics, agriculture, pathogen surveillance, and many other fields [60]. By enhancing our ability to detect, interpret, and predict genetic variation, these methods contribute broadly to scientific progress and generate knowledge that benefits both research and society.

With the rapid advancement of artificial intelligence and the expansion of technology, machine learning algorithms that achieve excellent results have become commonplace. Some of them operate like black boxes, creating a gap in the explainability of their functioning, which somewhat limits understanding of the peculiarities frequently present in biological research. Within the concepts of machine learning, we propose not only a classifier and dimensionality reducer, but also a novel method and software that contributes to genomic region analysis with high explainability of its results, which considers three important steps: 1) pattern discovery, aligned with data preparation; 2) dimensionality reduction; 3) classification of these data. Although the initial development focused on distance-value comparisons using a fuzzy set approach, experimental results revealed performance limitations. Consequently, we transitioned to more robust tree-based classification methods (Random Forest) to improve the model's classification performance.

This explainability, combined with open-source development, improves the reproducibility of the method and enables tailored adjustments across a wide range of applications, including the identification of conserved regions, gene families, and regulatory motifs; metagenomic profiling; and supervised classification of individuals into specific subspecies or populations, among others.

All GREAC codes, user manual, toy model example and scripts are publicly available on GitHub at https://github.com/SALIPE/GREAC, including details on codes and scripts in the section "Supplemental Methods".

Supplementary information. Supplementary information for this study is available in Appendix A, providing data tables, results from additional performance runs, and experimental results for model parameter definition that could not be included in the main text. This supplementary information ensures scientific reproducibility and provides readers with a deeper understanding of the study and its practical applications.

Abbreviations

GREAC: Genomic Region Extraction and Classifier; DSP: Digital Signal Processing; EIIP: Electron–Ion Interaction Potential; FFT: Fast Fourier Transform; RFE: Recursive Feature Elimination; RRM: Resonant Recognition Model; SNP: Single-Nucleotide Polymorphism; SVM: Support Vector Machine; SARS-CoV-2: Severe Acute Respiratory Syndrome Coronavirus 2; DENV: Dengue Virus; HBV: Hepatitis B Virus; HIV: Human Immunodeficiency Virus; MPOX: Monkeypox Virus; NCBI: National Center for Biotechnology Information; WHO: World Health Organization; REPL: Read–Eval–Print Loop.

Declarations

Ethics approval and consent to participate

Not applicable. This study did not involve human participants, human data, or human tissue, nor did it involve animal experimentation. All viral genome sequences analysed in this work were retrieved from publicly available repositories.

Consent for publication

Not applicable. This manuscript does not contain data from any individual person.

Availability of data and materials

The GREAC software, source code, user manual, example scripts, a toy-model dataset and execution logs used in this study are publicly available in the GREAC GitHub repository [31]: https://github.com/SALIPE/GREAC.

The viral genome datasets analysed during the current study were obtained from previously published sources and public repositories. The SARS-CoV-2 dataset corresponds to the collection used in KEVOLVE [25]; the HIV-1 dataset was obtained from CASTOR-KRFE [12]; the DENV dataset was derived from GRAMEP [11]. The HBV sequences, representing six genotypes, were obtained from HBVdb [32], and the Monkeypox complete genomes from both lineages were retrieved from the NCBI Virus database. All datasets used in this study are available in the GREAC GitHub repository for reproducibility.

Competing interests

The authors declare that they have no competing interests.

Funding

This study was financed by Fundação Araucária (Grant numbers 035/2019 and 138/2021, and NAPI-Bioinformática #66.2021) and by CNPq (440412/2022-6 and 408312/2023-8). This study was supported by FAPESP for M.H.P-Z. (grant number 2025/10318-5). This research was further supported by FCT — Fundação para a Ciência e Tecnologia, I.P., through the following projects: CIMO, UID/00690/2025 (DOI: 10.54499/UID/00690/2025) and UID/PRR/00690/2025; CeDRI, UID/05757/2025 (DOI: 10.54499/UID/05757/2025) and UID/PRR/05757/2025 (DOI: 10.54499/UID/PRR/05757/2025); and SusTEC, LA/P/0007/2020 (DOI: 10.54499/LA/P/0007/2020). The funding bodies had no role in the design of the study, the collection, analysis, or interpretation of data, or the writing of the manuscript.

Authors' contributions

FBS conceived and designed the GREAC software, implemented the source code, performed the computational experiments and analysed the results. MHPZ contributed to the k-mer extraction methodology and to the integration with GRAMEP. DH and MAP contributed to the biological interpretation of the discriminative genomic regions and to the validation on viral datasets. CB and JR supervised the algorithmic design and contributed to the high-performance computing aspects of the implementation. FML supervised the project and contributed to the methodological design and manuscript writing. All authors read, revised and approved the final manuscript.

Acknowledgements

The authors thank the developers of CASTOR-KRFE, KEVOLVE and GRAMEP for making their tools and datasets publicly available, which enabled the comparative evaluation reported in this work. The authors also acknowledge HBVdb and the NCBI Virus database for providing the additional viral genome sequences used in this study.

Appendix A Supplementary Information

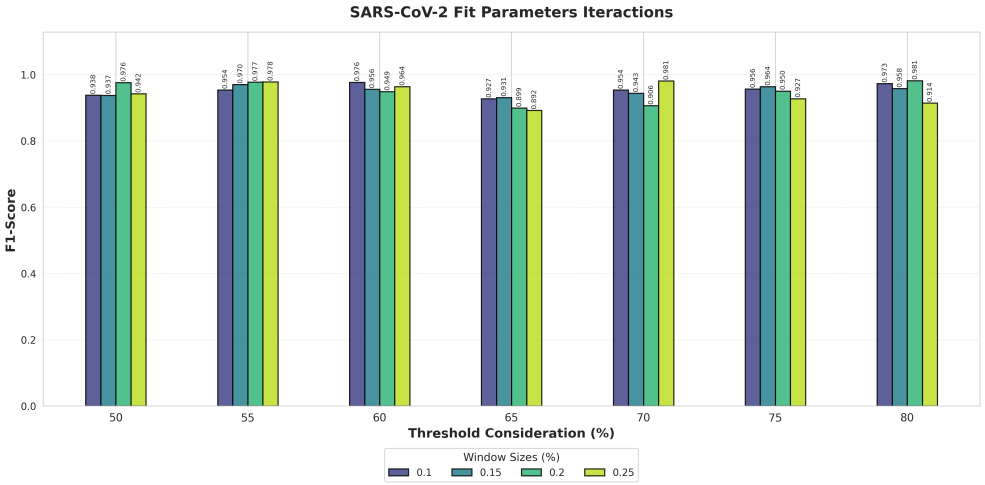


Fig. A1: Sars-CoV-2 fit parameters executions

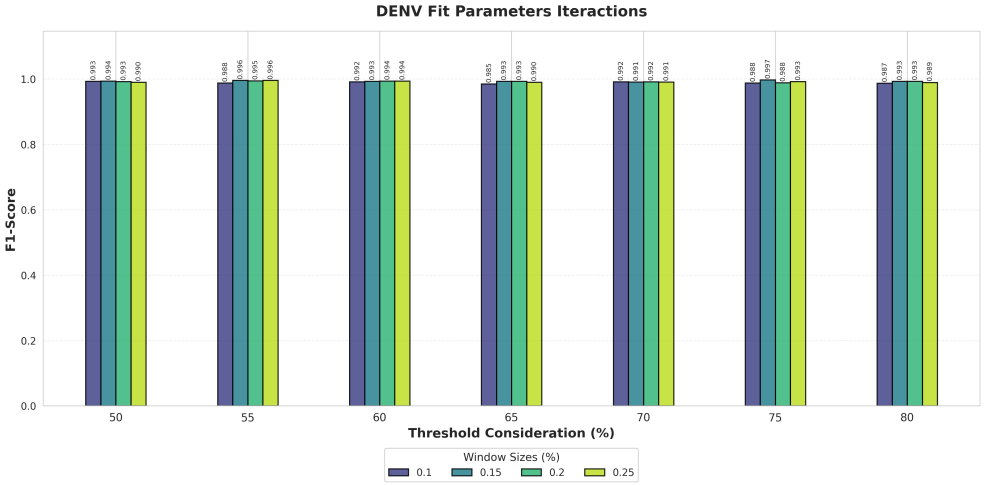


Fig. A2: DENV fit parameters executions

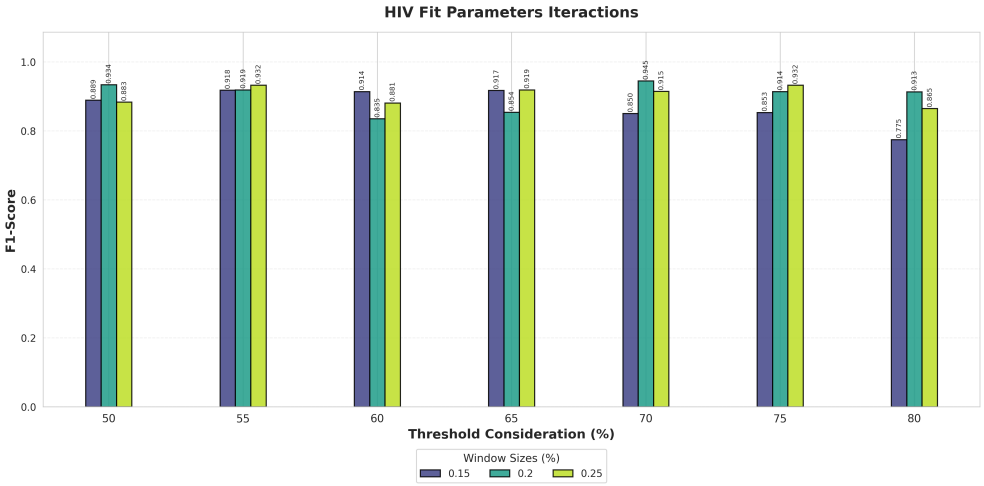


Fig. A3: HIV fit parameters executions

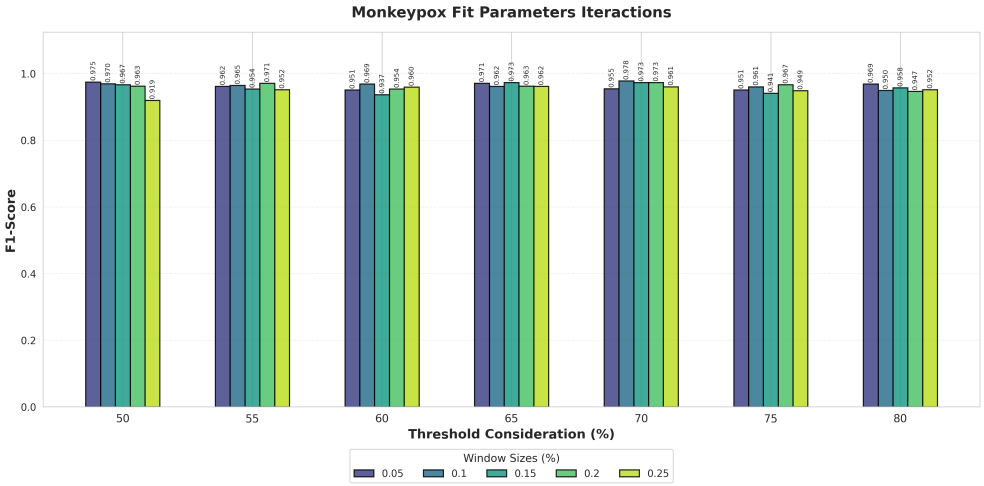
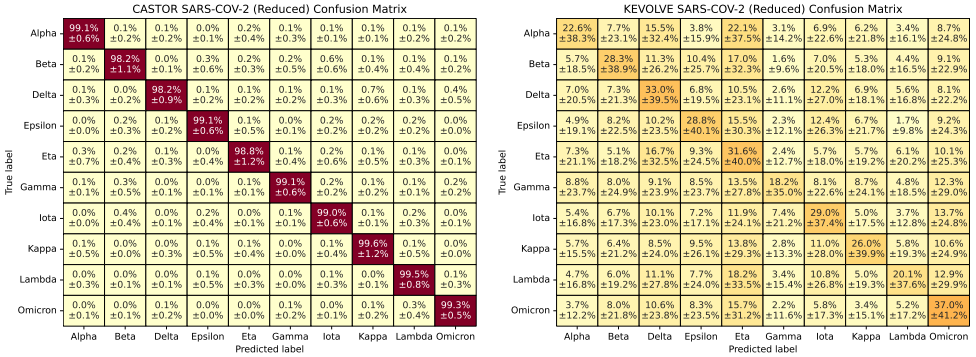


Fig. A4: Monkeypox fit parameters executions



(a) CASTOR-KRFE on the SARS-CoV-2 reduced dataset.

(b) KEVOLVE on the SARS-CoV-2 reduced dataset.

Fig. A13: Confusion matrix classification average accuracy and standard deviation over 100 iterations using the reduced dataset produced only with the regions extracted by GREAC.

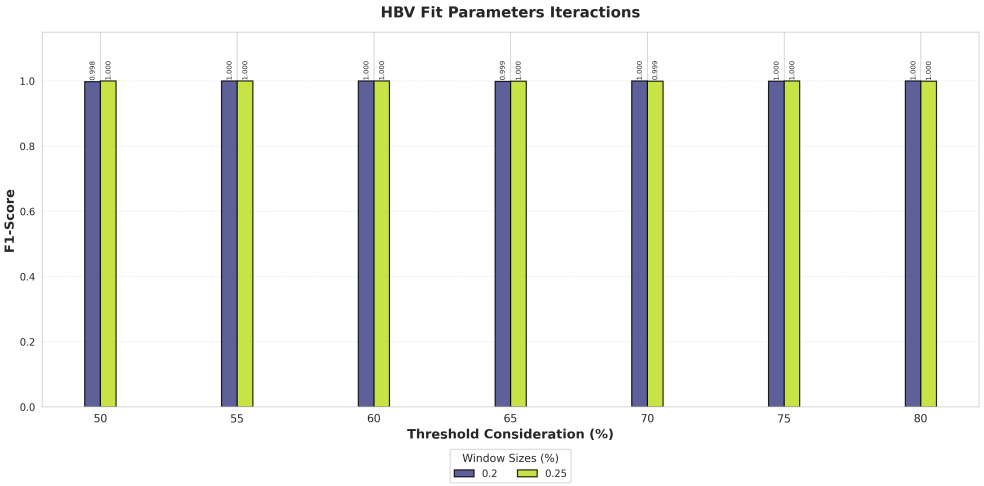
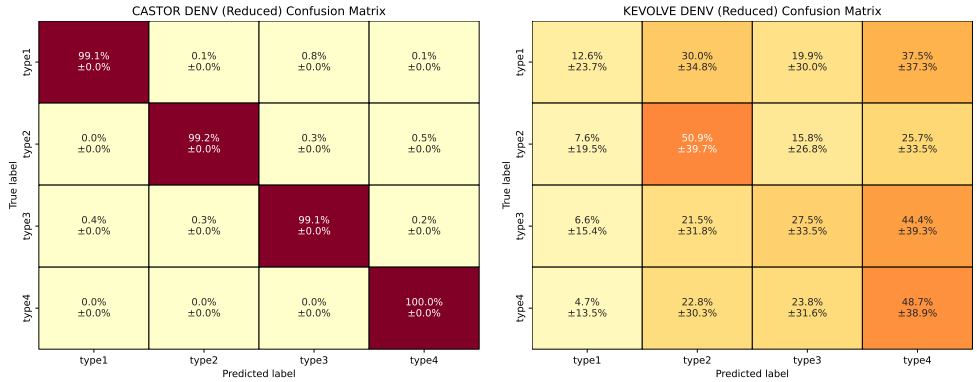


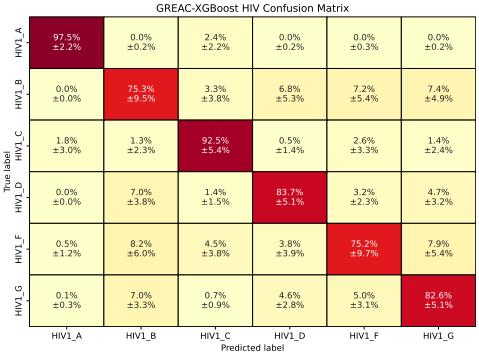
Fig. A5: HBV fit parameters executions



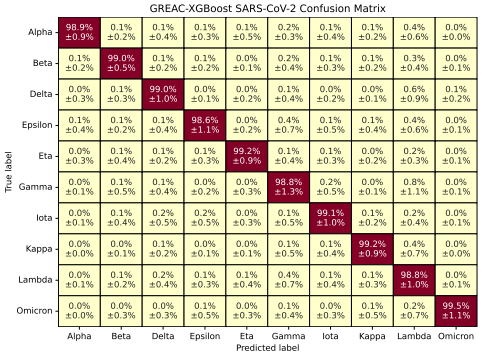
(a) CASTOR-KRFE on the DENV reduced dataset.

(b) KEVOLVE on the DENV reduced dataset.

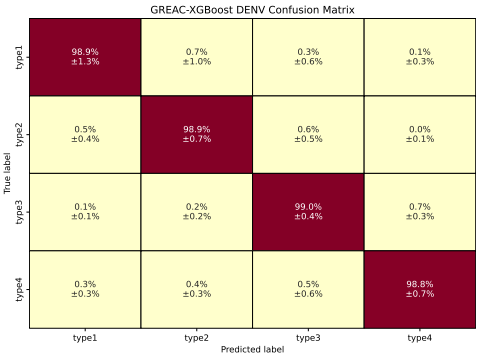
Fig. A14: Confusion matrix classification average accuracy and standard deviation over 100 iterations using the reduced dataset produced only with the regions extracted by GREAC.



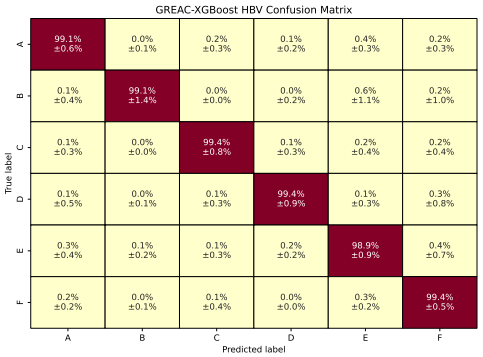
(a) HIV



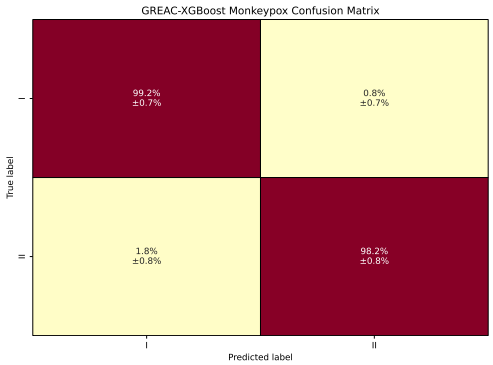
(b) SARS-CoV-2



(c) DENV

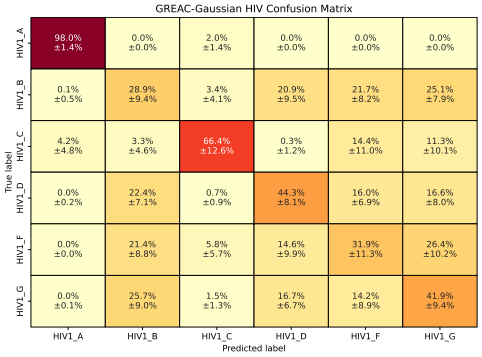


(d) HBV

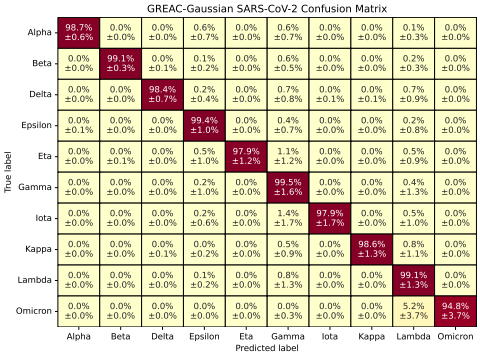


(e) Monkeypox

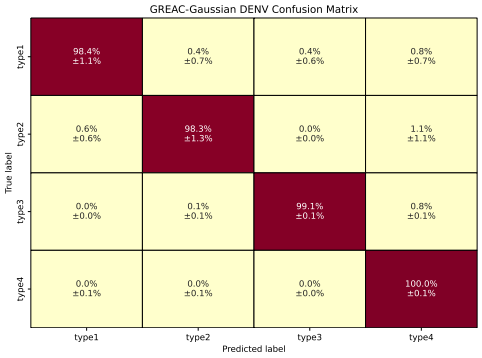
Fig. A6: Organisms XGBoost confusion matrix classification average accuracy percentage and standard deviation over 100 iterations. (A6a) HIV results. (A6b) SARS-CoV-2 results. (A6c) DENV results. (A6d) HBV results. (A6e) Monkeypox results.



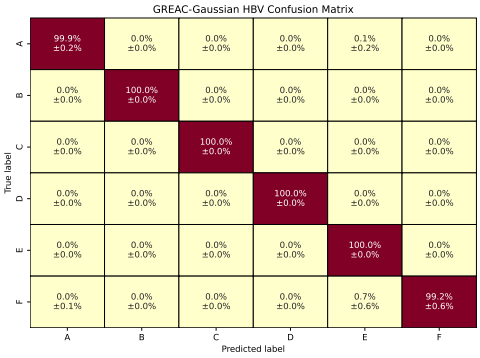
(a) HIV Gaussian Membership Results



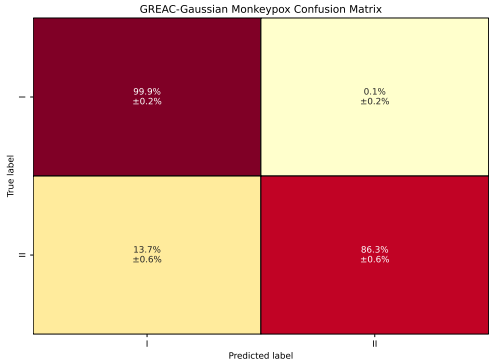
(b) SARS-CoV-2 Gaussian Membership Results



(c) DENV Gaussian Membership Results

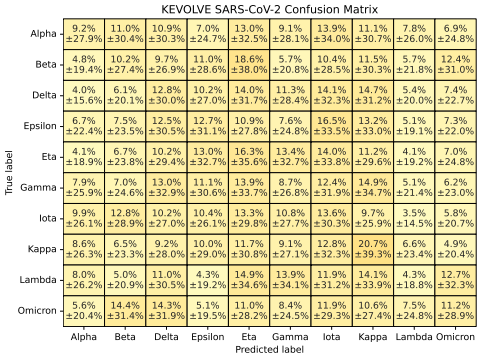


(d) HBV Gaussian Membership Results

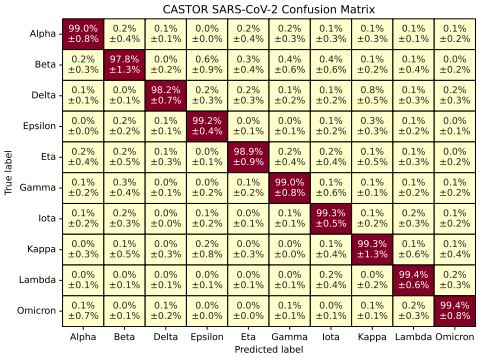


(e) Monkeypox Gaussian Membership Results

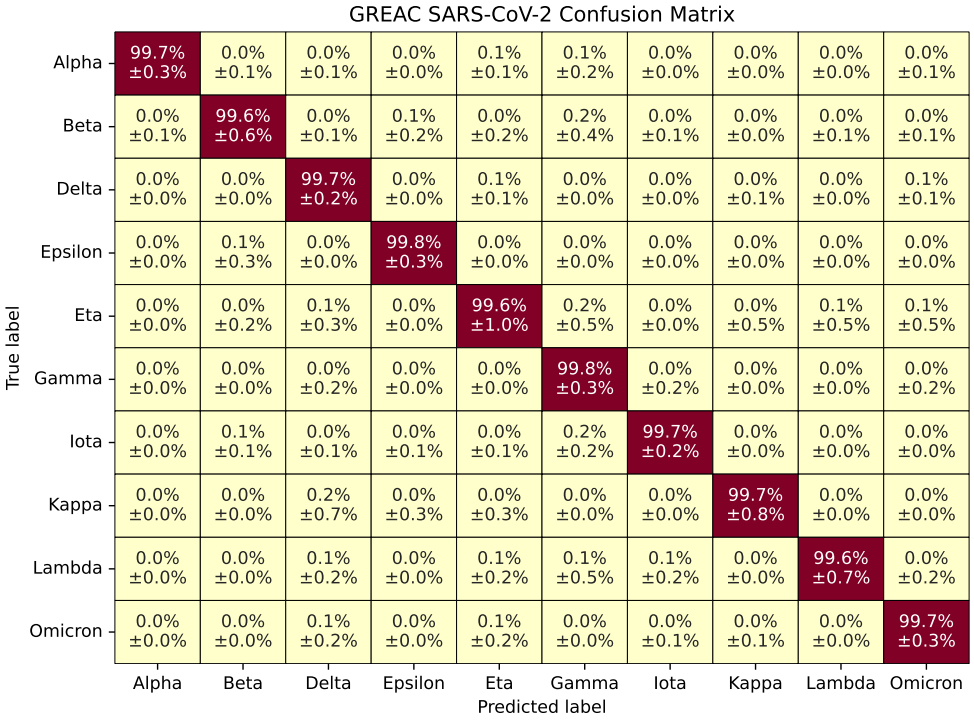
Fig. A7: Experiment using GRAMEP k-mer set using Gaussian Membership Classifications, confusion matrix results, average accuracy percentage, and standard deviation over 100 iterations.



(a) KEVOLVE

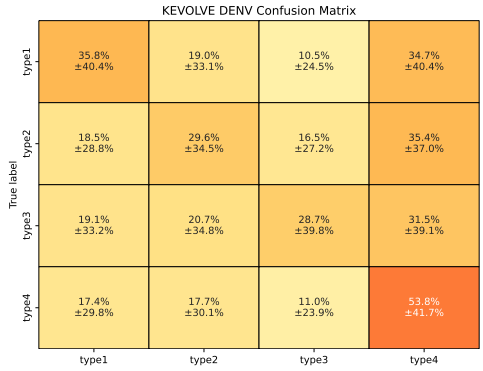


(b) CASTOR-KRFE

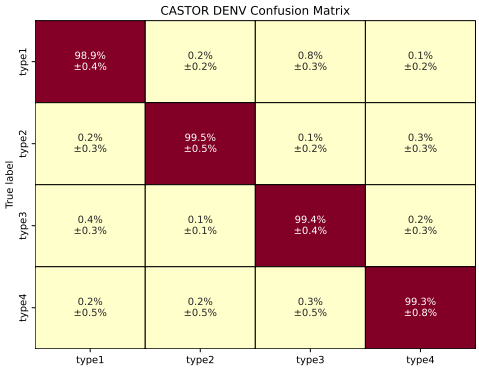


(c) GREAC

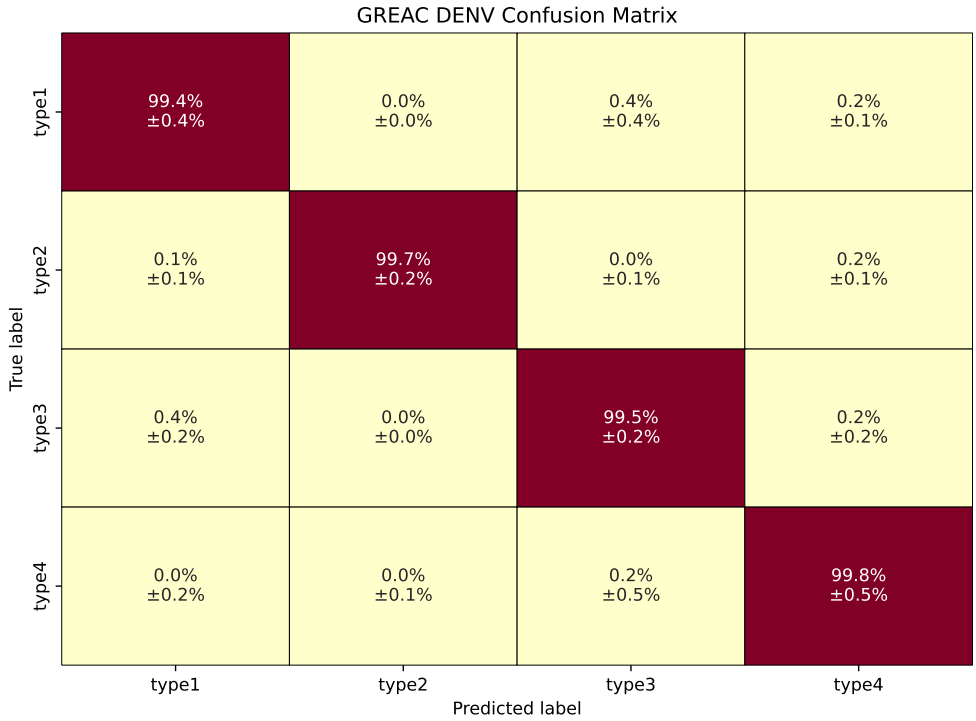
Fig. A8: SARS-CoV-2 confusion matrices, classifications, average accuracy percentage, and standard deviation over 100 iterations. (A8a) KEVOLVE results. (A8b) CASTOR-KRFE results.(A8c) GREAC results.



(a) KEVOLVE

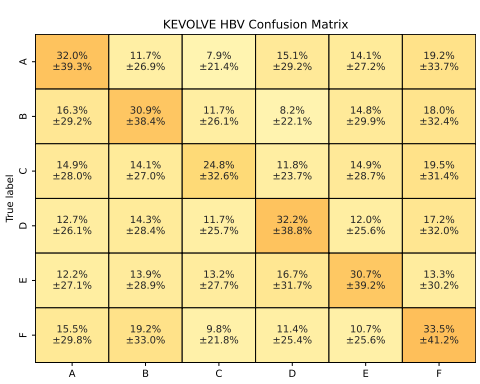


(b) CASTOR-KRFE

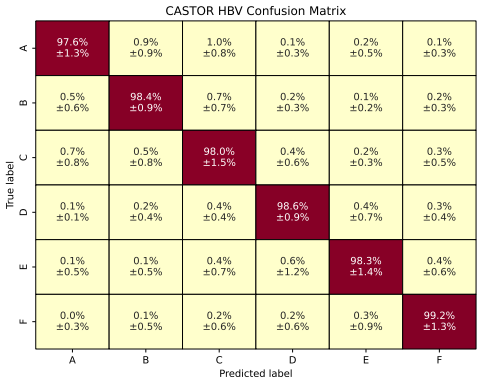


(c) GREAC

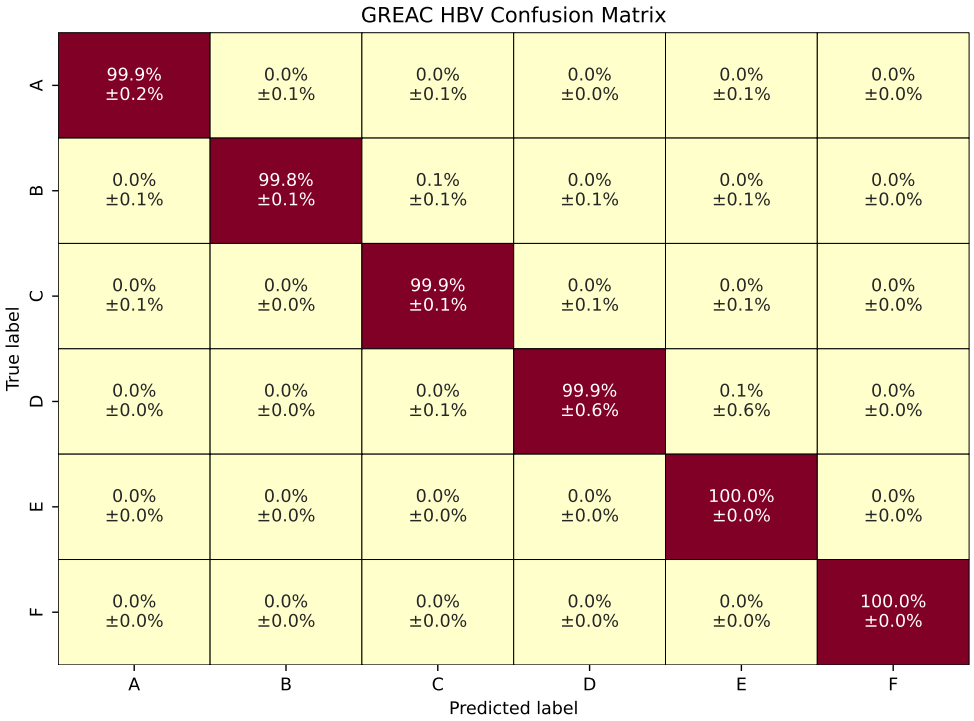
Fig. A9: DENV confusion matrices, classifications, average accuracy percentage, and standard deviation over 100 iterations. (A9a) KEVOLVE results. (A9b) CASTOR-KRFE results.(A9c) GREAC results.



(a) KEVOLVE

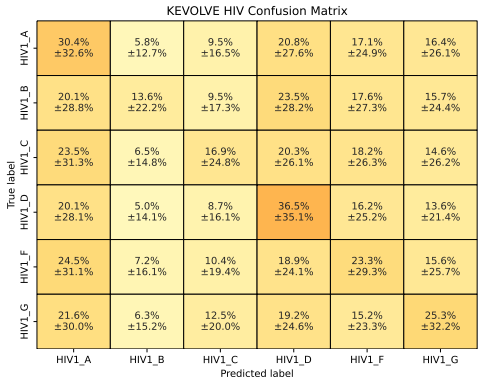


(b) CASTOR-KRFE

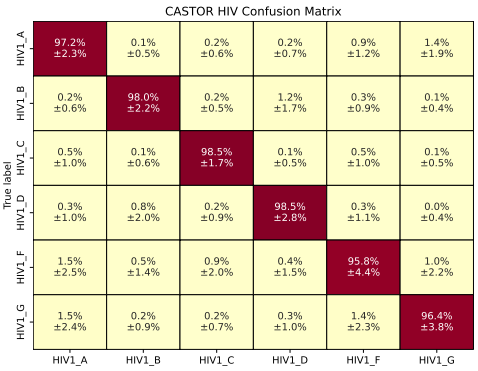


(c) GREAC

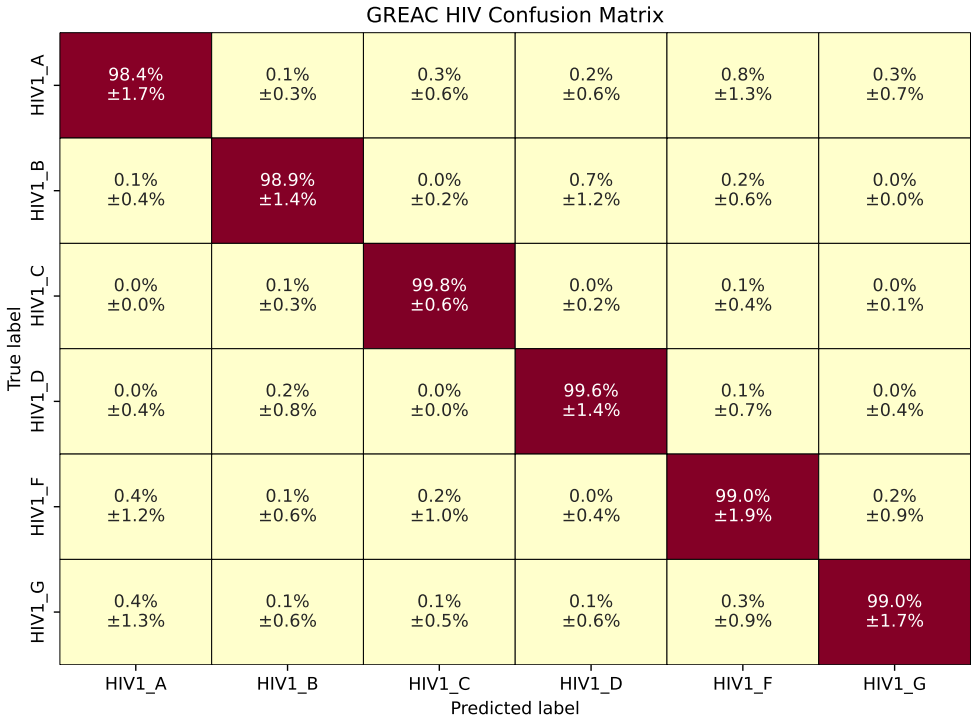
Fig. A10: HBV confusion matrices, classifications, average accuracy percentage, and standard deviation over 100 iterations. (A10a) KEVOLVE results. (A10b) CASTOR-KRFE results.(A10c) GREAC results.



(a) KEVOLVE

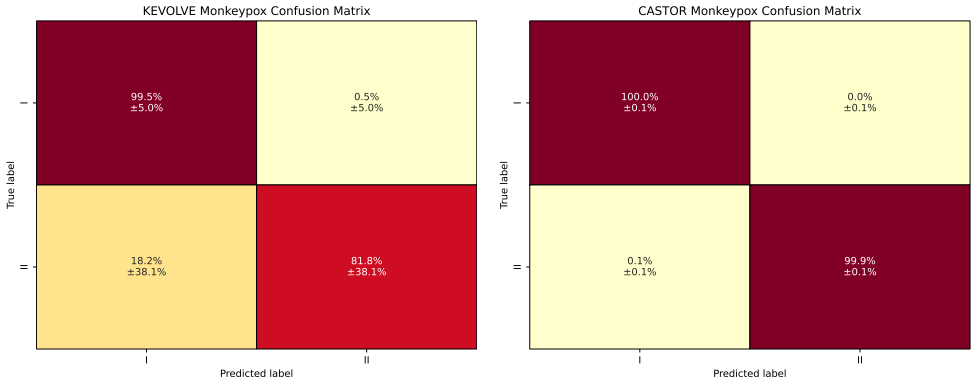


(b) CASTOR-KRFE



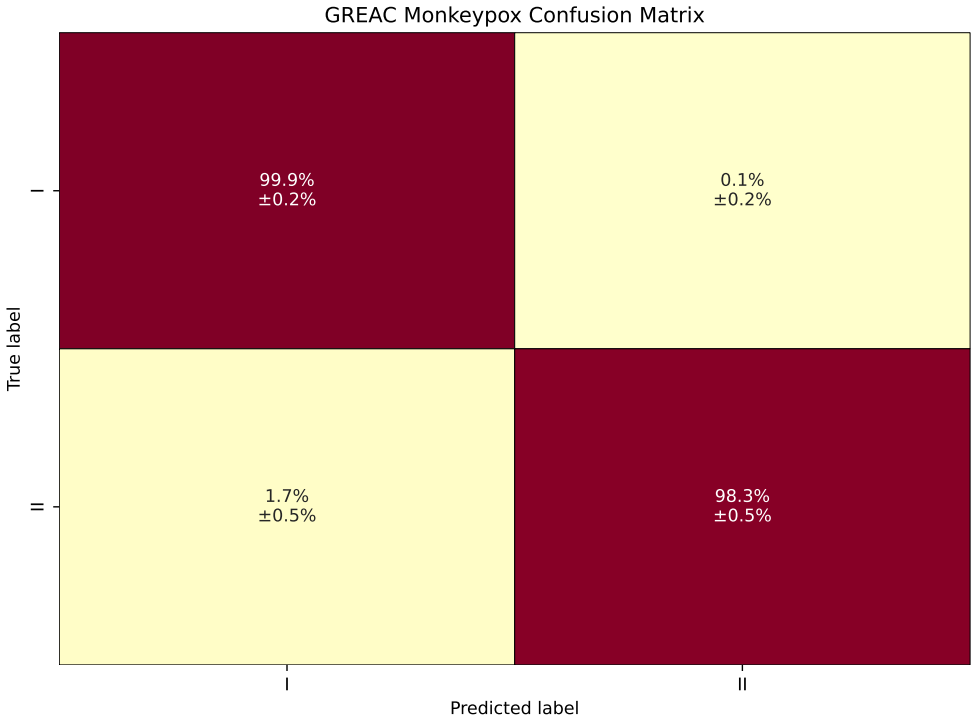
(c) GREAC

Fig. A11: HIV confusion matrices, classifications, average accuracy percentage, and standard deviation over 100 iterations. (A11a) KEVOLVE results. (A11b) CASTOR-KRFE results.(A11c) GREAC results.



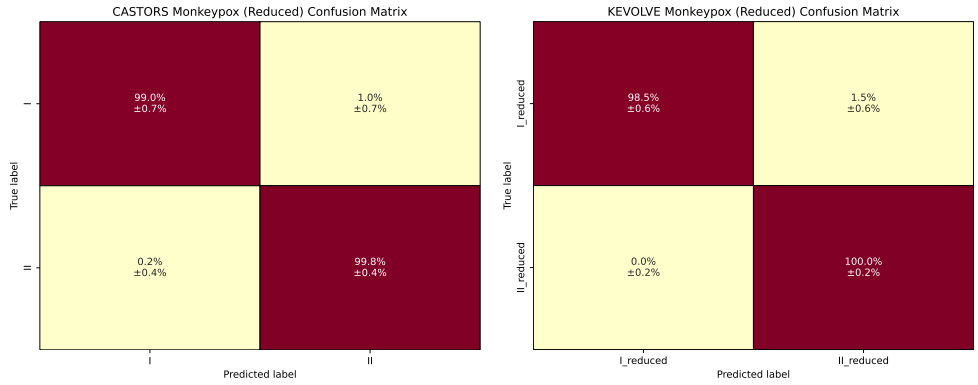
(a) KEVOLVE

(b) CASTOR-KRFE



(c) GREAC

Fig. A12: Monkeypox confusion matrices classifications, average accuracy percentage, and standard deviation over 100 iterations. (A12a) KEVOLVE results. (A12b) CASTOR-KRFE results.(A12c) GREAC results.



(a) CASTOR-KRFE results on Monkeypox reduced dataset.

(b) KEVOLVE results on Monkeypox reduced dataset.

Fig. A15: Confusion matrix classification average accuracy and standard deviation over 100 iterations using the reduced dataset produced only with the regions extracted by GREAC.

References

[1] Zielezinski, A., Vinga, S., Almeida, J., Karlowski, W.M.: Alignment-free sequence comparison: benefits, applications, and tools. Genome Biology **18**(1), 186 (2017) https://doi.org/10.1186/s13059-017-1319-7 . Accessed 2022-03-09

[2] Foxman, E.F., Iwasaki, A.: Genome–virome interactions: examining the role of common viral infections in complex disease. Nature Reviews Microbiology **9**(4), 254–264 (2011) https://doi.org/10.1038/nrmicro2541

[3] De Pierri, C.R., Voyceik, R., Mattos, L.G.C., Kulik, M.G., Camargo, J.O., Oliveira, A.M., Lima Nichio, B.T., Marchaukoski, J.N., Silva Filho, A.C., Guizelini, D., Ortega, J.M., Pedrosa, F.O., Raittz, R.T.: Sweep: representing large biological sequences datasets in compact vectors. Scientific Reports **10**(1) (2020) https://doi.org/10.1038/s41598-019-55627-4

[4] Perico, C.P., De Pierri, C.R., Neto, G.P., Fernandes, D.R., Pedrosa, F.O., Souza, E.M., Raittz, R.T.: Genomic landscape of the sars-cov-2 pandemic in brazil suggests an external p.1 variant origin. Frontiers in Microbiology **13** (2022) https://doi.org/10.3389/fmicb.2022.1037455

[5] Yamada, R., Okada, D., Wang, J., Basak, T., Koyama, S.: Interpretation of omics data analyses. Journal of human genetics **66**(1), 93–102 (2021)

[6] Berisha, V., Krantsevich, C., Hahn, P.R., Hahn, S., Dasarathy, G., Turaga, P., Liss, J.: Digital medicine and the curse of dimensionality. npj Digital Medicine **4**(1) (2021) https://doi.org/10.1038/s41746-021-00521-5

[7] Russell, S.J., Norvig, P., Davis, E.: Artificial Intelligence: A Modern Approach. Prentice Hall, Englewood Cliffs, NJ (2010)

[8] Lopes, F.M., Martins Jr, D.C., Cesar Jr, R.M.: Feature selection environment for genomic applications. BMC Bioinformatics **9**(1), 451 (2008)

[9] Barros-Carvalho, G.A., Van Sluys, M.-A., Lopes, F.M.: An efficient approach to explore and discriminate anomalous regions in bacterial genomes based on maximum entropy. Journal of Computational Biology **24**(11), 1125–1133 (2017)

[10] Breve, M.M., Pimenta-Zanon, M.H., Lopes, F.M.: BASiNETEntropy: an alignment-free method for classification of biological sequences through complex networks and entropy maximization (2022)

[11] Pimenta-Zanon, M.H., Kashiwabara, A.Y., Vanzela, A.L.L., Lopes, F.M.: Gramep: an alignment-free method based on the maximum entropy principle for identifying snps. BMC Bioinformatics **26**(1) (2025) https://doi.org/10.1186/s12859-025-06037-z

[12] Lebatteux, D., Remita, A.M., Diallo, A.B.: Toward an alignment-free method for feature extraction and accurate classification of viral sequences. Journal of Computational Biology **26**(6), 519–535 (2019)

[13] Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J.: Basic local alignment search tool. Journal of molecular biology **215**(3), 403–410 (1990)

[14] Zielezinski, A., *et al.*: Benchmarking of alignment-free sequence comparison methods. Genome Biology **20**(1), 144 (2019) https://doi.org/10.1186/s13059-019-1755-7 . Accessed 2023-05-27

[15] Vinga, S., Almeida, J.: Alignment-free sequence comparison—a review. Bioinformatics **19**(4), 513–523 (2003) https://doi.org/10.1093/bioinformatics/btg005

[16] Moeckel, C., Mareboina, M., Konnaris, M.A., Chan, C.S.Y., Mouratidis, I., Montgomery, A., Chantzi, N., Pavlopoulos, G.A., Georgakopoulos-Soares, I.: A survey of k-mer methods and applications in bioinformatics. Computational and Structural Biotechnology Journal **23**, 2289–2303 (2024) https://doi.org/10.1016/j.csbj.2024.05.025

[17] Kuksa, P., Pavlovic, V.: Efficient alignment-free dna barcode analytics. BMC bioinformatics **10**, 1–18 (2009) https://doi.org/10.1186/1471-2105-10-S14-S9

[18] Ito, E.A., Katahira, I., Vicente, F.F.d.R., Pereira, L.F.P., Lopes, F.M.: BASiNET - BiologicAl Sequences NETwork: a case study on coding and non-coding RNAs identification. Nucleic Acids Research **46**(16), 96–96 (2018) https://doi.org/10.1093/nar/gky462

[19] Bonidia, R.P., Sampaio, L.D.H., Lopes, F.M., Sanches, D.S.: Feature extraction of long non-coding rnas: A fourier and numerical mapping approach. In: Iberoamerican Congress on Pattern Recognition, pp. 469–479 (2019). Springer

[20] Souza, F., Pimenta-Zanon, M.H., Henriques, D., Pinto, M.A., Balsa, C., Rufino, J., Lopes, F.M.: Resonant recognition model as a preprocessing technique for rna classification. In: Guarda, T., Portela, F., Augusto, M.F. (eds.) Advanced Research in Technologies, Information, Innovation and Sustainability, pp. 3–17. Springer, Cham (2025)

[21] Haubold, B.: Alignment-free phylogenetics and population genetics. Briefings in Bioinformatics **15**(3), 407–418 (2013) https://doi.org/10.1093/bib/bbt083

[22] Van Etten, J., Stephens, T.G., Bhattacharya, D.: A k-mer-based approach for phylogenetic classification of taxa in environmental genomic data. Systematic Biology **72**(5), 1101–1118 (2023) https://doi.org/10.1093/sysbio/syad037

[23] Irber, L., Pierce-Ward, N.T., Abuelanin, M., Alexander, H., Anant, A., Barve, K., Baumler, C., Botvinnik, O., Brooks, P., Dsouza, D., Gautier, L., Hera, M.R.,

Houts, H.E., Johnson, L.K., Klötzl, F., Koslicki, D., Lim, M., Lim, R., Nelson, B., Ogasawara, I., Reiter, T., Scott, C., Sjödin, A., Standage, D., Swamidass, S.J., Tiffany, C., Vemuri, P., Young, E., Brown, C.T.: sourmash v4: A multitool to quickly search, compare, and analyze genomic and metagenomic data sets. Journal of Open Source Software **9**(98), 6830 (2024) https://doi.org/10.21105/joss.06830

[24] Chor, B., Horn, D., Goldman, N., Levy, Y., Massingham, T.: Genomic dna k-mer spectra: models and modalities. Genome biology **10**, 1–10 (2009) https://doi.org/10.1186/gb-2009-10-10-r108

[25] Lebatteux, D., Soudeyns, H., Boucoiran, I., Gantt, S., Diallo, A.B.: Machine learning-based approach kevolve efficiently identifies sars-cov-2 variant-specific genomic signatures. Plos one **19**(1), 0296627 (2024)

[26] Lebatteux, D., Soudeyns, H., Boucoiran, I., Gantt, S., Diallo, A.B.: Kanalyzer: a method to identify variations of discriminative k-mers in genomic sequences. In: 2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), pp. 757–762 (2022). https://doi.org/10.1109/BIBM55620.2022.9995370

[27] Solomatine, D.P., Ostfeld, A.: Data-driven modelling: some past experiences and new approaches. Journal of hydroinformatics **10**(1), 3–22 (2008)

[28] Duffy, S., Shackelton, L.A., Holmes, E.C.: Rates of evolutionary change in viruses: patterns and determinants. Nature Reviews Genetics **9**(4), 267–276 (2008)

[29] Peck, K.M., Lauring, A.S.: Complexities of viral mutation rates. Journal of Virology **92**(14) (2018) https://doi.org/10.1128/jvi.01031-17

[30] Patiño-Galindo, J.Á., Filip, I., Rabadan, R.: Global patterns of recombination across human viruses. Molecular Biology and Evolution **38**(6), 2520–2531 (2021) https://doi.org/10.1093/molbev/msab046

[31] Souza, F., GREAC Contributors: GREAC: Genomic Region Extraction and Classifier. https://github.com/SALIPE/GREAC (2025)

[32] Hayer, J., Jadeau, F., Deleage, G., Kay, A., Zoulim, F., Combet, C.: Hbvdb: a knowledge database for hepatitis b virus. Nucleic Acids Research **41**(D1), 566–570 (2012) https://doi.org/10.1093/nar/gks1022

[33] Bezanson, J., Edelman, A., Karpinski, S., Shah, V.B.: Julia: A fresh approach to numerical computing. SIAM Review **59**(1), 65–98 (2017)

[34] Wood, D.E., Salzberg, S.L.: Kraken: ultrafast metagenomic sequence classification using exact alignments. Genome Biology **15**(3) (2014) https://doi.org/10.1186/gb-2014-15-3-r46

[35] Rosen, G., Garbarine, E., Caseiro, D., Polikar, R., Sokhansanj, B.: Metagenome fragment classification using n-mer frequency profiles. Advances in Bioinformatics **2008**(1) (2008) https://doi.org/10.1155/2008/205969

[36] Breve, M.M., Lopes, F.M.: A simplified complex network-based approach to mrna and ncrna transcript classification. In: Setubal, J.C., Silva, W.M. (eds.) Advances in Bioinformatics and Computational Biology, pp. 192–203. Springer, Cham (2020)

[37] Cosic, I.: Macromolecular bioactivity: Is it resonant interaction between macromolecules? - theory and applications. IEEE transactions on bio-medical engineering **41**, 1101–14 (1995) https://doi.org/10.1109/10.335859

[38] Cosic, I.: The Resonant Recognition Model of Macromolecular Bioactivity: Theory and Applications. Birkhäuser, Basel, Switzerland (2012)

[39] Li, A., Zhang, J., Zhou, Z.: Plek: a tool for predicting long non-coding rnas and messenger rnas based on an improved k-mer scheme. BMC Bioinformatics **15**(1), 311 (2014) https://doi.org/10.1186/1471-2105-15-311

[40] Kang, Y.-J., Yang, D.-C., Kong, L., Hou, M., Meng, Y.-Q., Wei, L., Gao, G.: CPC2: a fast and accurate coding potential calculator based on sequence intrinsic features. Nucleic Acids Research **45**(W1), 12–16 (2017) https://doi.org/10.1093/nar/gkx428

[41] Kaplan, D.: Bayesian Statistics for the Social Sciences. Guilford Publications, New York, NY (2023)

[42] Duda, R.O., Hart, P.E., *et al.*: Pattern Classification. John Wiley & Sons, New York, NY. (2006)

[43] Zadeh, L.A.: Fuzzy sets. Information and control **8**(3), 338–353 (1965)

[44] Reddy, C.S., KVSVN, R.: An improved fuzzy approach for cocomo's effort estimation using gaussian membership function. J. Softw. **4**(5), 452–459 (2009)

[45] Wu, T.-J., Huang, Y.-H., Li, L.-A.: Optimal word sizes for dissimilarity measures and estimation of the degree of dissimilarity between dna sequences. Bioinformatics **21**(22), 4125–4132 (2005) https://doi.org/10.1093/bioinformatics/bti658

[46] Wu, T.-J., Burke, J.P., Davison, D.B.: A measure of dna sequence dissimilarity based on mahalanobis distance between frequencies of words. Biometrics, 1431–1439 (1997) https://doi.org/10.2307/2533509

[47] Kurtz, S., Narechania, A., Stein, J.C., Ware, D.: A new method to compute k-mer frequencies and its application to annotate large repetitive plant genomes. BMC genomics **9**, 1–18 (2008)

[48] Chen, T., Guestrin, C.: Xgboost: A scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. KDD '16, pp. 785–794. Association for Computing Machinery, New York, NY, USA (2016). https://doi.org/10.1145/2939672.2939785 . https://doi.org/10.1145/2939672.2939785

[49] Broder, A.Z.: On the resemblance and containment of documents. In: Proceedings. Compression and Complexity of SEQUENCES 1997 (Cat. No. 97TB100171), pp. 21–29 (1997). IEEE

[50] Ondov, B.D., Treangen, T.J., Melsted, P., Mallonee, A.B., Bergman, N.H., Koren, S., Phillippy, A.M.: Mash: fast genome and metagenome distance estimation using minhash. Genome Biology **17**(1) (2016) https://doi.org/10.1186/s13059-016-0997-x

[51] Paris, G., Robilliard, D., Fonlupt, C.: Exploring overfitting in genetic programming. In: Liardet, P., Collet, P., Fonlupt, C., Lutton, E., Schoenauer, M. (eds.) Artificial Evolution, pp. 267–277. Springer, Berlin, Heidelberg (2004)

[52] Rostami, M., Berahmand, K., Forouzandeh, S.: A novel community detection based genetic algorithm for feature selection. Journal of Big Data **8**(1), 2 (2021) https://doi.org/10.1186/s40537-020-00398-3

[53] Domingos, P.: The role of occam's razor in knowledge discovery. Data mining and knowledge discovery **3**, 409–425 (1999) https://doi.org/10.1023/A:1009868929893

[54] Majumdar, P., Niyogi, S.: Sars-cov-2 mutations: the biological trackway towards viral fitness. Epidemiology and Infection **149** (2021) https://doi.org/10.1017/s0950268821001060

[55] Miller, R.H., Kaneko, S., Chung, C.T., Girones, R., Purcell, R.H.: Compact organization of the hepatitis b virus genome. Hepatology **9**(2), 322–327 (1989) https://doi.org/10.1002/hep.1840090226

[56] Holmes, E.C.: The evolutionary biology of dengue virus. In: New Treatment Strategies for Dengue and Other Flaviviral Diseases: Novartis Foundation Symposium 277, pp. 177–192 (2006). Wiley Online Library

[57] Plummer, E., Buck, M.D., Sanchez, M., Greenbaum, J.A., Turner, J., Grewal, R., Klose, B., Sampath, A., Warfield, K.L., Peters, B., Ramstedt, U., Shresta, S.: Dengue virus evolution under a host-targeted antiviral. Journal of Virology **89**(10), 5592–5601 (2015) https://doi.org/10.1128/jvi.00028-15

[58] Aksamentov, I., Roemer, C., Hodcroft, E., Neher, R.: Nextclade: clade assignment, mutation calling and quality control for viral genomes. Journal of Open Source Software **6**(67), 3773 (2021) https://doi.org/10.21105/joss.03773

[59] Barbieri, M.: What is code biology? Biosystems **164**, 1–10 (2018) https://doi.org/10.1016/j.biosystems.2017.10.005

[60] Moeckel, C., Mareboina, M., Konnaris, M.A., Chan, C.S.Y., Mouratidis, I., Montgomery, A., Chantzi, N., Pavlopoulos, G.A., Georgakopoulos-Soares, I.: A survey of k-mer methods and applications in bioinformatics. Computational and Structural Biotechnology Journal **23**, 2289–2303 (2024) https://doi.org/10.1016/j.csbj.2024.05.025

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- GREACsupplementary.pdf