IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 6 August 2023, accepted 29 August 2023, date of publication 4 September 2023, date of current version 8 September 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3311752

RESEARCH ARTICLE

Discovering Mathematical Patterns Behind HIV-1 Genetic Recombination: A New Methodology to Identify Viral Features

ANA GUERRERO-TAMAYO[1], BORJA SANZ URQUIJO[1], CONCEPCIÓN CASADO[2], MARÍA-DOLORES MORAGUES TOSANTOS[3], ISABEL OLIVARES[2], AND IKER PASTOR-LÓPEZ[1]

[1]Faculty of Engineering, University of Deusto, Bilbao, 48007 Biscay, Spain
[2]National Microbiology Center (NMC), Instituto de Salud Carlos III (ISCIII), Majadahonda, 80523 Madrid, Spain
[3]Faculty of Medicine and Nursing, University of the Basque Country (UPV/EHU), Leioa, 48940 Biscay, Spain

Corresponding author: Ana Guerrero-Tamayo (ana.guerrero@deusto.es)

This work was supported by the Research Training Grants Program, University of Deusto.

ABSTRACT In this article, we introduce a novel methodology for characterizing viral genetic features: the Unified Methodology of recombinant virus Identification (UMI). Our methodology converts genomic sequences into spectrograms, applies transfer learning using a pre-trained Convolutional Neural Network (CNN), and employs interpretability tools to identify the genomic regions relevant for characterizing a viral sequence as recombinant. The UMI methodology does not necessitate multiple sequence alignment or manual adjustments. As a result, it operates much faster, has low computational demands, and is capable of handling substantial amounts of data. To validate this, we applied UMI to one extensively studied and documented case: HIV-1 genetic recombination. We worked with all identified HIV-1 complete sequences (13554 sequences up to 2020), searching for mathematical patterns, signatures, that characterize an HIV-1 sequence as recombinant. CNN's hit rate (test accuracy) is 94%, with consistent and differentiated decision areas in each category. Using interpretability tools, we verified that the hot zones were similar for sequences of the same subtype and phylogenetic proximity. The leading areas for classifying a sequence as recombinant or non-recombinant are coincident with genomic regions that play a key role in genetic recombination processes. By applying UMI methodology we found that there is indeed a genome mathematical pattern that assesses an HIV-1 sequence as recombinant. In addition, we located its position. Considering expert knowledge, our results showed a substantial, robust and biologically-consistent hit rate. This type of solution can successfully guide the location and subsequent characterization of relevant areas, avoiding the heavy analysis of multiple sequence alignment and manual adjustments.

INDEX TERMS Convolutional neural network, deep learning, genetic recombination, genome mathematical pattern, genome mathematical signature, HIV-1.

I. INTRODUCTION

The capacity of viruses to vary severely complicates health care and prevention policies [1]. Mutation is one of the main mechanisms underlying viral diversity. It consists of introducing random errors during the replication process. These errors may result in differences in some viral characteristics [2]. Genetic recombination is another frequent phenomenon. This happens when two different strains of a virus, or two different viruses, co-infect a cell at the same time and the result of this replication is a new hybrid genome, called recombinant [3]. Recombinant viruses are usually life-incompatible. However, in some cases, they can produce new pandemic viruses. This is the case of SARS-CoV2, MERS-CoV, Influenza Virus or Human Immunodeficiency Virus Type 1 (HIV-1) [4].

Recombination in RNA viruses is an exchange of genetic information between genomic RNA molecules [5], which occurs during viral replication usually by a mechanism

The associate editor coordinating the review of this manuscript and approving it for publication was Kathiravan Srinivasan.

(a) HIV-1 env gp160 3D Protein (b) HIV-1 env gp160 FASTA File

FIGURE 1. 3D Graphical Representation of a Protein and its Coding Genomic Sequence. Fig. 1a shows the three-dimensional structure of the protein gp160 encoded by the *env* gene. Q79670 (ENV_HV1MV) Human immunodeficiency virus type 1 group O (isolate MVP5180) (HIV-1) Envelope glycoprotein gp160. See https://swissmodel.expasy.org/. Fig. 1b shows the FASTA file containing the genomic sequence of the HIV-1 *env* gene, complete cds (ID. L42371).

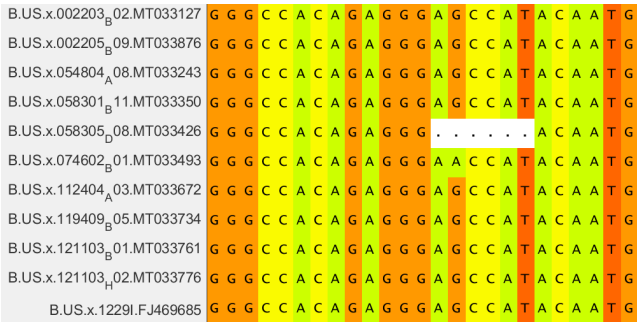


FIGURE 2. HIV-1 Multiple Sequence Alignment. By introducing GAPS (·), multiple sequence alignment algorithms search the better position for each nucleotide of each sequence so a direct comparison is possible. After applying these algorithms, experts usually must make manual adjustments to reach the optimal alignment. This figure was generated using MATLAB2021b Computational Biology Apps.

known as copy-choice which consists of a switch of the growing RNA from the genomic RNA template of a virus to the genomic RNA of another, resulting in a hybrid viral genome containing RNA fragments from both parental genomes [6]. Recombinant viruses have been detected with high frequency in various RNA viruses, such as picornaviruses [7] and coronaviruses [8]. But until now, HIV-1 is one of the viruses with the most recombinant sequences characterized, and a large amount of information on whole genome sequences is found in sequence databases.

The genome contains all the genetic information of an organism, coded in a sequence of four nucleotides (A, C, G, and T). It includes protein coding regions, regulatory regions and additional elements [9]. Fig. 1 shows an example of a genomic sequence (HIV-1 *env* gene) and its encoded protein.

Viruses resulting from genetic recombination can cause pandemics with new and potentially devastating effects [10], [11], [12], [13]. The HIV-1 pandemic caused a dramatic social impact in the 1980s and, consequently, high funding for the development of the AIDS cure. Despite all the efforts, HIV-1 variability severely affects the current lack of effective treatments and vaccines (genetic recombination is one of the determinant changeability factors). Therefore, it is necessary to understand the process of genetic recombination and identify all the possible mechanisms that determine its occurrence and its outcome. The importance of genetic recombination in viral diversification [14] renders the development of new methods to detect recombinant viruses very significant, and the search for mathematical patterns (inherent mathematical regularities and structures in the DNA sequence). Applying emerging technologies, such as Deep Learning, can help improve the current level of microbiological knowledge with new approaches. HIV-1 is an excellent study model because of its high number of identified and labelled recombinant sequences.

Traditional methodologies for the analysis and comparison of genome sequences are based on multiple sequence alignment. This requires introducing gaps to place each nucleotide, letter by letter, in the same position and then make a direct comparison of the aligned sequences (Fig. 2). There are several multiple sequence alignment algorithms, including Clustal [15], TCoffee [16], Probcons [17], MAFFT [18], Muscle [19], MSAprobs [20], and GLProbs [21]. These algorithms are complex, computationally expensive, and do not allow the handling of large amounts of data; therefore, they provide sub-optimal solutions [22].

There is still no efficient solution for the analysis of large numbers of sequences and no objective method has been approved by the entire scientific community. The best approaches are to apply several algorithms, select their common results, and then perform manual adjustments based on expert microbiological knowledge.

We propose a new methodology using a genomic representation independent of sequence length, and the implementation of Deep Learning algorithms to model the categorization of recombinant sequences [23] with interpretability tools that point to the genomic areas involved in genetic recombination. The Unified Methodology of recombinant virus Identification, UMI Methodology (see Section III-D), achieves excellent and robust results, with the added complexity that both recombinant and non-recombinant sequences also undergo mutations. Thus, the complexity is even greater. We applied several interpretability tools to identify where the CNN relies on to determine the classification of each sample.

- This article outlines the detection of mathematical patterns, mathematical signatures, embedded in the genome that characterize the Human Immunodeficiency Virus Type 1 (HIV-1) as recombinant.

- UMI allows not only the detection of the recombinant feature of each sample but also targets the genome areas where

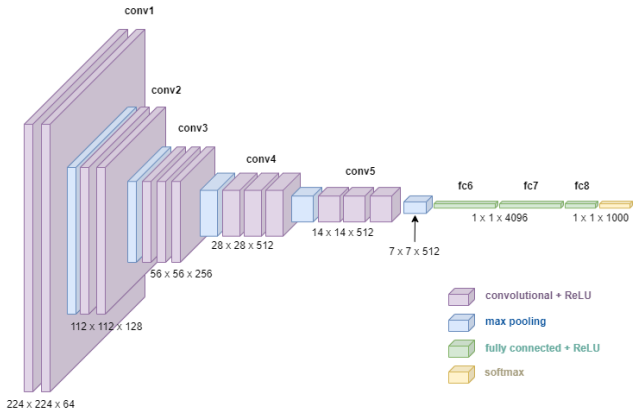


FIGURE 3. Pre-trained Convolutional Neural Network VGG-16 Architecture Schematic. Once the image "enters" the CNN, convolution phase begins. This consists of taking groups of nearby pixels from the input image and transforming them mathematically (scalar product) to filter different features. Pooling phase emphasizes the most important features detected by each convolutional filter. The output of pooling phase is the pooled feature map (numerical matrix). Next step converts this feature matrix to a vertical vector, which is the input to the fully connected layer. Its neurons detect certain features, then communicate their values to softmax neurons and they check out each feature and decide about its relevancy. Finally, each output neuron gives a classification result and a confidence score (how sure this neuron is about its decision). This figure is based on the diagram drawn by Kenneth Leung, which illustrates VGG-16 architecture [30].

this characterization is located. This will help to gain more knowledge about the processes of genetic recombination in HIV-1.

The remainder of this paper is organized as follows. Section II describes the background of the UMI methodology. Section III contains a complete description of the designed method. Section IV details the results. Section V (Biological Consistency Analysis) delves into the peculiarities of the HIV-1 replication cycle and the microbiological meaning of the obtained results, with multiple examples of coherence between them and expert microbiological knowledge. This document ends with Conclusions, summarizing the microbiological consistency of our results, and Future Studies.

II. BACKGROUND

Convolutional Neural Networks (CNN) have demonstrated high efficacy in detecting patterns in images not visible to the human eye [24], [25], [26], [27]. They are even applied in the detection of HIV outbreaks using genetic data [28]. Fig. 3 depicts the architecture of one of the most widely used CNNs, VGG-16 [29].

The optimal use of this potential requires biologically consistent graphical conversion of genomic sequences [31]. Genome frequency-domain analysis is a promising graphical representation for detecting biological patterns [32], [33], [34].

The application of signal processing tools to genomic sequences of different organisms has shed light on the relationship between nucleotide periodicity and various features [35], such as the relationship between coding areas and sensitivity at frequency $f = 1/3$ [36], non-coding RNA molecules location [37], GpC islands detection or microsatellites identification [38].

In addition, these tools have the advantage of not requiring multiple sequence alignments to compare different genomes. This reduces the computational resources needed and eliminates the useless information of aligned sequences [39].

Specifically, spectrogram imaging, as a graphical representation of the Fast Fourier Transform (FFT), is increasingly used to analyze genomic sequences [40]. The first step in generating the spectrogram of a sequence was to convert that sequence into four digital signals (U_a, U_g, U_c, U_t), which are activated by the presence of a nucleotide type (A, G, C, T) at each position.

$$U_\alpha(x_i) = \begin{cases} 1, & \text{if } x_i = \alpha \\ 0, & \text{otherwise} \end{cases} \tag{1}$$

The next step was to calculate the spectrogram for each of the four binary signals obtained from this decomposition.

The Fast Fourier Transform (FFT) corresponding to each of these signals was performed, and this is a computationally efficient method for the calculation of the Discrete Fourier Transform (DFT) [41].

$$X(k) = \sum_{n=0}^{N-1} U_\alpha(x_i) W_N^{kn} \quad 0 \le k \le N-1 \tag{2}$$

where:

$$W_N = e^{-j2\pi/N} \tag{3}$$

FFT algorithms decompose an N-point DFT into smaller DFTs [42] using sliding windows.

Although the application of Deep Learning Tools in health is still controversial [43], multiple studies have demonstrated that the performance of Deep Learning Tools using spectrograms renders good results in the interpretation of medical results [44], [45], [46], [47], [48], [49].

Calculation of the spectrogram of a genomic sequence makes it possible to obtain a graphic representation of the significance of each periodicity (frequency) at each position (assimilating the concept of time to position) of the genome.

The x-axis represents the position of each nucleotide. The y-axis is the frequency range and the z-axis is the value of the FFT. The spectrogram is a two dimensional representation obtained by replacing the z-axis with a color palette according to the amplitude of the FFT of the position of a nucleotide (x-axis) at a given frequency (y-axis) [50].

Applying Transfer Learning on a pre-trained CNN allows extending its potential to new types of images and new domains such as medical diagnostics: prediction of multiple diseases from chest X-ray images [51], classification of abdominal ultrasound images [52], diabetic retinopathy images [53], cellular morphological changes to identify cellular phenotypes (incipient cancer process and so on) [54] or detection of Major Depressive Disorder (MDD) by observing

electroencephalogram (EEG) signals transformed to spectrograms [55].

Therefore, the application of pre-trained CNNs as a tool for the classification of genomic sequence spectrograms can yield interesting and positive results. In this case, the HIV-1 compete sequences were classified as recombinant or non-recombinant.

The spectrogram is a visual tool that represents the spectral composition of a signal as a function of time using the Fast Fourier Transform (FFT). This image contains a significant mathematical component.

The identification of invisible mathematical patterns that determine the recombination of a sequence can potentially offer increased robustness compared to traditional algorithms used for detecting genetic recombination events in nucleotide sequences. These traditional methods rely on statistical approaches, which may raise concerns regarding their performance and reliability [56].

Interpretability tools can provide information on areas that are determinants of subsequent classification [57], [58], [59]. The results of the interpretability analysis make it possible to evaluate which factors determine whether an HIV-1 sequence is recombinant or not.

Three different interpretability tools can identify different areas of interest, and the results can be combined to obtain the maximum information:

1) Grad-CAM (Gradient-weighted Class Activation Mapping). It uses gradients of the classification score with respect to the final convolutional feature map [60]. Grad-CAM is intuitive, but its resolution is low.

2) LIME (Local Interpretable Model-agnostic Explanations). It computes feature importance by segmenting an image into several features and generating ''artificial'' observations by randomly including or excluding these features [61], [62]. Its accuracy can vary but it exhibits good visual interpretability.

3) Gradient Attribution. It accurately identifies the specific pixels in the image that are crucial for the decision-making process of the Deep Learning tool. It returns a pixel map that highlights the most influential pixels on the class score when changed [63], [64]. Its accuracy is very high; however, visually, it is very noisy.

In terms of accuracy and ease of use for the human eye, the advantages and disadvantages of these three tools make them particularly interesting to work with simultaneously, enabling the comparison of results and obtaining complementary data on features and patterns.

III. MATERIALS AND METHODS

All experiments ran in this equipment:

- Processing Unit: Intel(R) Xeon(R) Gold 5220R CPU @ 2.20 GHz.
- Installed RAM: 256 GB usable.
- Operative System: Windows 10 Education. Version: 21H1.
- GPU: NVIDIA RTX A6000. Total memory: 179.451 GB. VRAM: 48.571 GB.

TABLE 1. Dataset Structure. Non-Recombinants account for 82.36% of the total (13554 complete HIV-1 sequences until 2020). Recombinants are the 17.64%. Percentages in parentheses are calculated with respect to the total number of sequences of training, validation and test sets (11267, 1987 and 300 sequences).

Dataset	Recombinant	Non-Recombinant	TOTAL
Train	1903 (16.89%)	9364 (83.11%)	11267
Validation	338 (17.01%)	1649 (82.99%)	1987
Test	150 (50.00%)	150 (50.00%)	300
TOTAL	2391	11163	13554

A. HIV-1 COMPLETE GENOMIC SEQUENCES COMPENDIUM

HIV-1 genome sequences were downloaded from the Los Alamos National Laboratory HIV Database (https://www.hiv.lanl.gov/), all subtypes, complete genome. Recombinant sequences are named as Circulating Recombinant Forms (CRFs). The entire genome of these viruses has been demonstrated to exhibit genetic recombination. This means that certain regions of the genome cluster with one subtype, whereas other regions cluster with a different subtype. To facilitate the quick identification of subtypes, CRF expression is omitted, referring to recombinants only by acronyms after the hyphen symbol. Therefore, CRF29-BF1 is referred to as BF1.

To construct the dataset, we labelled all CRF sequences as recombinant and non-CRFs as non-recombinant. Subsequently, we randomly distributed these sequences into training, validation, and test folders according to the guidelines outlined in Table 1.

As shown in Table 1, the test dataset ponderation (50 % / 50 %) deliberately differs from that corresponding to train and validation sets (17 % / 83 %) to detect under-fitting biases during the training phases [65].

B. SPECTROGRAM GENERATION

Working with two different graphical representations of the spectrogram of a sequence allows us to obtain complementary information that may clarify the possible pattern that determines whether an HIV-1 genome is recombinant.

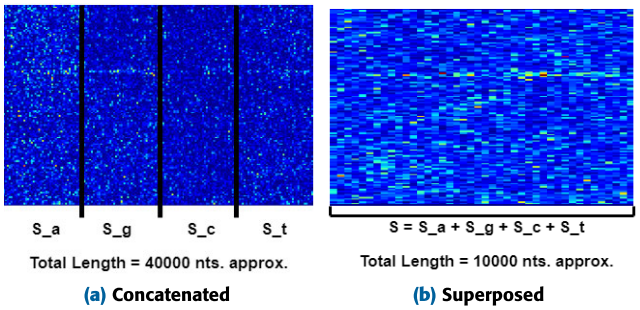
The first one, called **Concatenated Spectrogram**, displays the concatenation of the spectrogram for each of the four types of nucleotides on the x-axis. A schematic is shown in Fig. 4a.

This representation system can determine the most influential nucleotide types for the characterization of a virus as recombinant. However, it is difficult to determine accurately the involved area of the genome.

The second one is called **Superposed Spectrogram**. In this representation, the z-axis is the arithmetic sum of the values along the z-axis for each of the four nucleotide types:

$$S = S_a + S_g + S_c + S_t \tag{4}$$

Fig. 4b shows the graphical concept of this type of representation, which allows us to more effectively specify the influential zones of the genome.



(a) Concatenated (b) Superposed

FIGURE 4. Spectrograms Schemes. The x-axis represents the position of each nucleotide. Therefore, in Fig. 4a the length is four times the sequence length owing to the concatenation of the four spectrograms. In Fig. 4b the length of x-axis matches the genome length. In both cases, the y-axis represents a frequency range from 0 to 0.5. The z-axis is the spectrogram calculation.

FIGURE 5. Jet Color Map. Blue color corresponds to lowest sensitivities and red to the highest ones. See jet colormap at https://matplotlib.org.

To work with two-dimensional spectrograms, the z-axis is assimilated to a "jet" color palette. See Fig. 5. Using the matplotlib.pyplot.pcolormesh function, we mapped the minimum value of the FFT to deep blue (0) and the maximum value to deep red (1), automatically generating color assignments for each intermediate value of the FFT. By default, the data range is mapped to the colorbar range using linear scaling.

Each spectrogram is identified by the reference of the corresponding sequence.

We generated spectrograms of both datasets using Python, the Scipy library, scipy.signal.spectrogram. To avoid biases and bad practices during CNN training, we omitted axes, margins and any other elements that may distort the operation of the network. Therefore, the image is limited to the spectrogram only [66], [67]. We performed all the experiments by splitting the entire dataset into training, validation and test folders. The complete dataset is available in the Supplementary Information - Dataset folder. CNNs trained with the same information in concatenated and superposed. This improves the comparability of results by eliminating GAPs related to different learning data.

C. PRE-TRAINED CNN SELECTION CRITERIA

Once both spectrogram datasets were generated, the next step was to perform a test bench to select not only the CNN with the best performance, but also the set of hyperparameters that achieved the best and most balanced accuracy levels between non-recombinant and recombinant viral genomes.

CNNs of choice were:

- VGG16 [68].
- ResNET - 101 [69].
- Inception V3 [70].

The experiments were performed using the MATLAB2021b App Deep Learning Designer. These CNNs are pre-trained with millions of images from the ImageNet database, and they can classify images into 1000 categories. More information about the pre-trained convolutional neural networks in MATLAB is available at https://www.mathworks.com.

Through Transfer Learning, the CNN learns the specific features of both datasets. Freezing the weights of all intermediate layers intact, we replaced the last fully connected layer, specifying the new number of categories to analyze. We also replaced the classification layer of the pre-trained model. In this way, we extract all previous knowledge from the other datasets and adapt it to our new problem.

As demonstrated by preliminary tests, Data Augmentation degraded the performance significantly, probably because it altered the biological meaning of the sequence. Applying this technique would be counterproductive, particularly if rotations are generated. This is because the biological meaning of the image is determined by its exact orientation. Therefore, we did not apply this technique.

The following test bench was performed to standardize the tests on concatenated and superposed spectrogram datasets with the most direct comparability possible between different CNNs (Table 2).

TABLE 2. Test Bench. It was applied under the same conditions to all the selected CNNs. It consists of different combinations of Learning Rate (0.0001 and 0.0100), Batchsize (52 and 128) and Epochs (10 and 30).

Test	Learning Rate	Batchsize	Epochs
1)	0.0001	52	10
2)	0.0100	52	10
3)	0.0001	52	30
4)	0.0100	52	30
5)	0.0001	128	10
6)	0.0100	128	10
7)	0.0001	128	30
8)	0.0100	128	30

1) BEST PERFORMANCE BASED DECISION CRITERIA

The criteria for determining the best results should not be solely based on the total number of hits. The four selected performance metrics of the CNN are as follows:

- Validation Accuracy.
- Training Time.
- Confusion Matrix Results.
- AUC.

Training Time is a measure of the required computation. For similar Validation Accuracy and AUC values, the decisive criterion was the maximum Test Accuracy in both recombinant and non-recombinant categories (Confusion Matrix Results). The best performance was achieved when these hit rates in both categories were balanced and optimal.

According to this criterion, the CNN that performs best in both datasets (concatenated and superposed) was the one selected. The balance should not only exist in both categories, but also in both datasets.

The results of all experiments and training curves are available at the Supplementary Information - Training Curves folder.

2) ROBUSTNESS CRITERIA

To determine the robustness of the results, the same optimized training with exactly the same dataset was repeated three times. If the results of these three attempts are similar, the system is considered to be robust.

3) INTERPRETABILITY TOOLS' RESULT GENERATION

To locate the positions of these mathematical patterns that enable the CNN to classify a genomic spectrogram as recombinant, we applied MATLAB functions such as gradCAM, imageLIME, and gradientMap.

These three functions generate a scoremap, which is a numerical matrix highlighting the relevant areas that contribute to the classification task.

These hot zones are output in image format. To achieve this, they overlaid a color map onto the original image, making the hot regions clearly visible.

In the case of Grad-CAM and LIME, a 256-color Jet Color Map is applied (see Fig. 5), from 0 (deep blue) to 1 (deep red). Gradient Attribution provides a pixel-resolution map that shows the most relevant pixels to the network's classification. This map is scaled with 255 colors from 0 (white) to 1 (black).

The hot maps of the complete dataset are available at the Supplementary Information - Hot Zones folder.

4) PHYLOGENETIC TREE AND COMPARISON WITH INTERPRETABILITY TOOLS' PATTERNS

A comparison was made between sequences with similar interpretability patterns and the phylogenetic tree of the 13554 HIV-1 complete sequences [71], [72]. First, we downloaded the HIV-1 complete consensus multiple sequence alignment from the Los Alamos National Laboratory HIV-1 Database (until 2020). These alignments contain a single sequence per patient [73]. Thus, we avoid biases due to comprehensive data from a few sources and to make the resultant phylogenetic tree more user-friendly.

An analysis was then performed between the similarity level of the patterns indicated by the interpretability tools, and the position of the sequences in the phylogenetic tree (the closeness between them). A phylogenetic tree of these multiple sequence alignments and phylogenetic distances between two sequences (Path Length), were calculated using MATLAB2021b - Computational Biology Apps.

When the analyzed sequence did not appear in the HIV-1 consensus multialignment, the representative sequence from the same patient was taken as a reference. Sometimes, it was not possible to locate the position of some sequences in the phylogenetic tree, because alignments lacked either the sequence itself or the patient's reference sequence. The patient was not identified in some sequences.

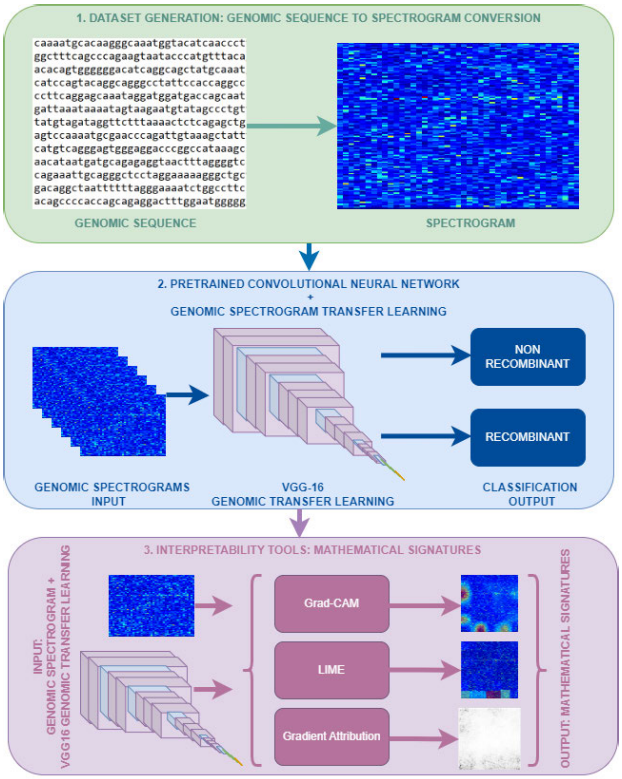


FIGURE 6. UMI Methodology Block Diagram. The first phase is a conversion of genomic sequences to biologically-coherent images, this is, genomic spectrograms. The second phase is applying transfer learning to a pre-trained Convolutional Neural Network (CNN) with this dataset of frequency-domain images, so the CNN can detect features in this type of images. Once the CNN is trained, interpretability tools allow to identify which areas the CNN looks at to classify each image. The result of the third and final phase is the location of mathematical signatures related with a viral feature.

The traceability of intra-patient sequences (virus sequences produced in the same patient's organism) was determined by the Patient Code (ID) field.

We loaded the HIV-1 complete consensus multiple sequence alignment into the Sequence Alignment Tool. We automatically generated the phylogenetic tree using the "View Tree" option. The phylogenetic path and distance are displayed by selecting two sequences on the tree itself.

This phylogenetic tree is available at the Supplementary Information - Phylogenetic Tree folder.

D. UMI METHODOLODY BLOCK DIAGRAM

Our new methodology, the Unified Methodology of recombinant virus Identification (UMI), consists of three phases, as described in Fig. 6.

IV. RESULTS

A. CNN SELECTION

To improve the reading and understanding of Confusion Matrix Results in the following sections, only the number of hits and failures are shown, according to Fig. 7).

TABLE 3. Best Results Abstract. Once all test combinations were performed, and considering the selection criteria, the best performance and the best hyperparameters for each of the pre-trained CNNs are shown.

Concatenated Spectrogram Dataset							
CNN	Hyperparameters	Validation Accuracy	Training Time	Training Time per Epoch	Test Accuracy	Confusion Matrix Results	AUC
VGG-16	Learning Rate: 0.0001 Batchsize: 52 Epochs: 10	94.92%	17 min 36 sec	1 min 45.60 sec	94.00%	145 – 5 13 – 137	0.9772
Resnet-101	Learning Rate: 0.0100 Batchsize: 52 Epochs: 10	94.82%	41 min 7 sec	4 min 6.70 sec	94.33%	149 – 1 16 – 134	0.9928
Inception-V3	Learning Rate: 0.0100 Batchsize: 52 Epochs: 30	94.82%	148 min 1 sec	4 min 56.03 sec	96.33%	146 – 4 7 – 143	0.9943
Superposed Spectrogram Dataset							
CNN	Hyperparameters	Validation Accuracy	Training Time	Training Time per Epoch	Test Accuracy	Confusion Matrix Results	AUC
VGG-16	Learning Rate: 0.0001 Batchsize: 52 Epochs: 10	92.90%	15 min 59 sec	1 min 35.90 sec	94.00%	143 – 7 11 – 139	0.9888
Resnet-101	Learning Rate: 0.0100 Batchsize: 52 Epochs: 30	96.83%	136 min 24 sec	4 min 32.80 sec	92.67%	143 – 7 15 – 135	0.9899
Inception-V3	Learning Rate: 0.0100 Batchsize: 52 Epochs: 30	96.48%	149 min 6 sec	4 min 58.20 sec	95.00%	149 – 1 14 – 136	0.9879

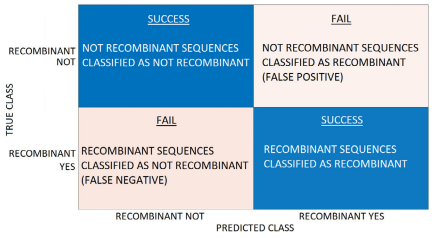


FIGURE 7. Confusion Matrix Scheme. The top row refers to non-recombinants. Hits are on the left and misses area on the right. The bottom row refers to recombinants. Failures are on the left and hits area on the right.

As shown in Table 3, both for concatenated and superposed spectrograms:

VGG-16 achieved better performances with a Learning Rate = 0.0001. The optimal Learning Rate of ResNET-101 and Inception V3 is 0.0100.

All the three CNNs improved with Batchsize = 52 vs. Batchsize = 128.

These three CNNs exhibited the same behavior as the number of Epochs increased. The success rate increased directly proportional to the number of Epochs classifying non-recombinant sequences, achieving 100%. However, for recombinant sequences, the success rate increased until it reaches its maximum, after which it began to decrease.

VGG-16 yielded better results with Epochs = 10 vs. Epochs = 30. ResNET-101 performed better with Epochs = 10 when working with concatenated spectrograms. However, if it worked with superposed spectrograms, Epochs = 30 was a better choice. For both datasets, Inception V3 slightly improved its performance for Epochs = 30.

VGG-16 training time was the lowest. Inception V3 developed the slowest learning process.

Because Res-NET 101 obtained the lowest accuracy for both datasets, it was discarded. Inception V3 provided better results for concatenated spectrograms but its accuracy was slightly lower for superposed recombinant spectrograms.

Therefore, in accordance with the performance measurement parameters defined in Section III-C1, VGG-16 was the CNN selected for this research. It yielded more balanced hit rates on the test dataset between non-recombinant and recombinant categories, and its training times were substantially lower. The best success rate and the most balanced results were achieved with 10 Epochs.

The best VGG-16 configuration for both concatenated and superposed spectrograms was:

- Learning Rate = 0.0001
- Batchsize = 52
- Epochs = 10

B. ROBUSTNESS

We performed the robustness test in accordance with the specifications described in Section III-C2. The same experiment, which was performed three times, achieved very similar results. Thus, the robustness of the method was demonstrated.

The results were similar in all three cases for both datasets (Tables 4 and 5); thus, the performance was robust. The most successful training was selected according to the criteria outlined in Section III-C1. In both datasets, Test 1 achieved the best hit rates and moreover the best balance between recombinants and non-recombinants.

C. HOT ZONES DETERMINATION

Interpretability tools allow us to identify where a CNN determines the recombinant character of a sequence. These spectrogram areas are called hot zones (where the CNN looks at to classify a sequence).

Figs. 8 and 9 represent the most repeated interpretability hot zones. In order to extract the most recurrent hot zones with the highest level of accuracy, we analyzed the interpretability results of each sequence in the test set individually (a total of 300 sequences). For the test set sequences that the network correctly predicted, we identified the most frequently occurring hot areas as well as the secondary ones. For the sequences where the network failed, we determined the most frequently occurring erroneous hot zones as well as the secondary ones.

The Grad-CAM and LIME results, although less accurate, easily render hot zones to the naked eye. So, they are more

TABLE 4. VGG-16 Concatenated Spectrograms Three Identical Test Results. These tests were performed with the best resulting configuration: Learning Rate= 0.0001, Batchsize=52 and NºEpochs=10.

Concatenated Spectrogram Dataset					
	Validation Accuracy	Training Time	Test Accuracy	Confusion Matrix Results	AUC
Test 1	94.92%	17 min 36 sec	94.00%	145 – 5 13 – 137	0.9772
Test 2	95.82%	15 min 51 sec	92.67%	144 – 6 16 – 134	0.9828
Test 3	95.02%	17 min 36 sec	90.33%	142 – 8 21 – 129	0.9714

TABLE 5. VGG-16 Superposed Spectrograms Three Identical Test Results. These tests were performed with the best resulting configuration (same as concatenated): Learning Rate= 0.0001, Batchsize=52 and NºEpochs=10.

Superposed Spectrogram Dataset					
	Validation Accuracy	Training Time	Test Accuracy	Confusion Matrix Results	AUC
Test 1	92.90%	15 min 59 sec	94.00%	143 – 7 11 – 139	0.9888
Test 2	92.25%	16 min 9 sec	93.33%	144 – 6 14 – 136	0.9904
Test 3	94.46%	16 min 39 sec	93.33%	144 – 6 14 – 136	0.9872

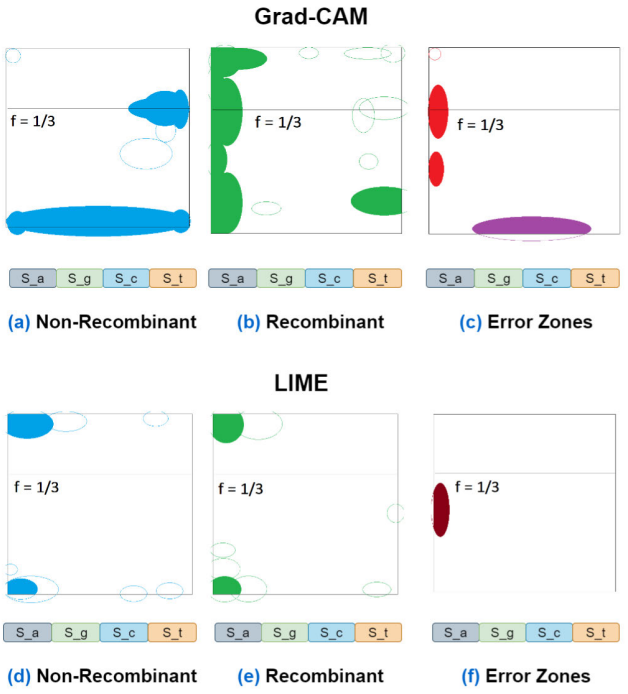


FIGURE 8. Concatenated Hot Zones. The most common hot areas are color-filled. The not filled ones represent hot areas with lower repeatability. In the case of Grad-CAM,(a) shows the hot areas where the CNN looked at to classify a sequence as non-recombinant. (b) shows those corresponding to the recombinant category. (c) indicates where did the CNN look at in misclassification cases. "Recombinant classified as non-recombinant" hot zones are purple-colored and "Non-recombinant classified as recombinant" hot zones are red-colored. In the case of LIME, (d) shows the hot areas where the CNN looked at to classify a sequence as non-recombinant. (e) shows those corresponding to the recombinant category. When the CNN fails, the LIME hot zones are indicated in (f) in brown.

suitable for the first approximation to the zones of interest. By contrast, Gradient Attribution provides a difficult visual interpretation.

Grad-CAM hot zones were analyzed, as shown in Fig. 8a., the low-frequency range of the four nucleotides plays an important role in determining a sequence as non-recombinant, as well as the frequencies close to the one-third for the T nucleotide. In recombinant concatenated spectrograms (Fig. 8b.), the leading Grad-CAM hot zone pivots around the left border (A nucleotide area) and in the lower third of the right border (T nucleotide area). There is a zone of interest common to non-recombinants and recombinants, located on the right border at the height of one-third of the T-nucleotide area. In the misclassification cases, the CNN looked at areas corresponding to the opposite category. The most common error zones are as shown in Fig. 8c.

In the case of LIME applied to concatenated spectrograms (8d. and 8e.), LIME targets the left corners (A nucleotide high and low frequencies) both for non-recombinant and recombinant. In several failure cases, the CNN was fixed in places that did not coincide with hot zones. This error zone is located around the middle of the left axis (average A nucleotide frequencies), as shown in 8f.

For non-recombinant superposed spectrograms, Grad-CAM hot zones are shown in Fig. 9a. The leading one corresponds to the high and low-frequency areas, corresponding mainly to 5'LTR and 3'LTR (the four corners). Secondary hot zones are concentrated at the upper and lower edges and at high and low frequencies.

Analyzing recombinant superposed spectrograms, Grad-CAM main influential areas (Fig. 9b.) are located in low frequencies and one-third frequency of 5'LTR.

In non-recombinant superposed spectrograms, LIME is overwhelmingly fixed in the lower-left corner (low-frequency area at the beginning of the sequence, 5'LTR primarily). To a lesser extent, in the upper-left corner (high frequency at the beginning of the sequence) and in the lower-right corner

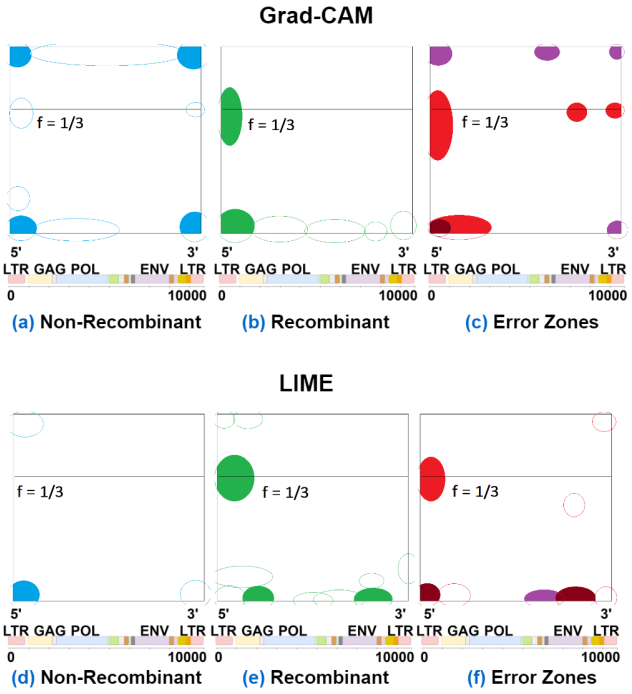


FIGURE 9. Superposed Hot Zones. The most common hot areas are color-filled. The not filled ones represent hot areas with lower repeatability. (a) shows the hot areas where the CNN looked at to classify a sequence as non-recombinant, and (d) presents the corresponding ones to LIME. (b) and (e) shows those corresponding to the recombinant category of Grad-CAM and LIME, respectively. (c) and (f) indicate where did the CNN look at in misclassification cases. "Recombinant classified as non-recombinant" hot zones are purple-colored and "Non-recombinant classified as recombinant" hot zones are red-colored. Brown-colored area refers to the common error zone between the two types of misclassification.

(low-frequency zone at the end of the sequence, 3'LTR). See Fig. 9d.

The significant regions in the superposed spectrograms of recombinant sequences are located at approximately one-third of the frequency range at the beginning of the sequence and some areas that approximate the *gag* and *env* genes. See Fig. 9e.

In superposed spectrograms, Grad-CAM and LIME indicate that the CNN misclassified looking at the hot zones of the opposite category. See Fig. 9c. and 9f.

The behaviors of the three interpretability tools are similar in concatenated and superposed spectrograms. Both methods obtain robust results and similar zones of interest in the nearby sequences.

V. BIOLOGICAL CONSISTENCY ANALYSIS

As explained in Section III-C4, after confirming that the CNN could differentiate between recombinant and non-recombinant sequences based on distinct decision patterns, the next step was to deepen the biological interpretation of this behavior. We generated the phylogenetic tree of all the complete HIV-1 sequences up to 2020. Next, we compared the similarity level of the hot zones with phylogenetic distance. We also compared hot zones with microbiological expert knowledge to validate the biological coherence or our results.

To facilitate the biological meaning of the hot areas, we provide a set of all the figures in Tables 6 to 11 with their scaled reference. See Supplementary Information -Hot Zones Reference Information.

A. HIV-1 MICROBIOLOGICAL CONSIDERATIONS

HIV-1 is a retrovirus, the genome of which contains two copies of single stranded RNA. The hallmark of retroviral replication is reverse transcription [74], that generates a linear DNA duplex from single stranded genomic RNA. Reverse transcription occurs via a series of steps, including two switches of nascent DNA, from one end to the opposite end of the RNA template [75]. As a result, two identical sequences called Long Terminal Repeats (LTRs), are formed at each end of DNA. After reverse transcription, the viral DNA is integrated into the genome of the host cell.

LTRs play a vital role in the initiation and regulation of viral transcription. Despite the identity of the sequence, each one has different functions. 5'LTR contains the signals for the initiation and regulation of viral transcription. 3'LTR performs in post-transcriptional processes. 3'LTR has an efficient polyadenylation site (adding a polyA tail to messenger RNA), and it is fundamental in the editing of viral transcripts [76], [77].

The two copies of contained RNA, as well as the switches occurring during reverse transcription, may be related to the high level of genetic recombination that occurs in an HIV infectious process [78]. Reverse transcriptase indistinctly uses portions of these two RNA molecules contained in each virion as a template [79].

HIV-1 is classified into four groups named M, N, O and P. Group M, which is the most widespread in the world, is divided into nine subtypes: A, B, C, D, F, G, H, J and K. Multiple recombinants between subtypes are currently circulating; they are called Circulating Recombinant Forms (CRF) [80].

B. ANALYSIS OF SUCCESSFUL CLASSIFICATIONS

The hot areas of sequences belonging to the same subtype are usually similar. In most cases, the closer the phylogenetic distance, the greater the similarity of the hot zones. A representative example is subtype 22-01A1 AY371165 and KF716462 (considering its reference sequence JN864058). See Table 6. The phylogenetic closeness between the two sequences is proven in Fig. 10. This path is highlighted in red.

The similarity level is higher for intra-patient sequences [81]. Therefore, interpretability tools are expected to yield even more similar results. As an example, Table 7 shows that there is a high similarity between the hot zones of concatenated and superposed spectrograms at sequences KR820306

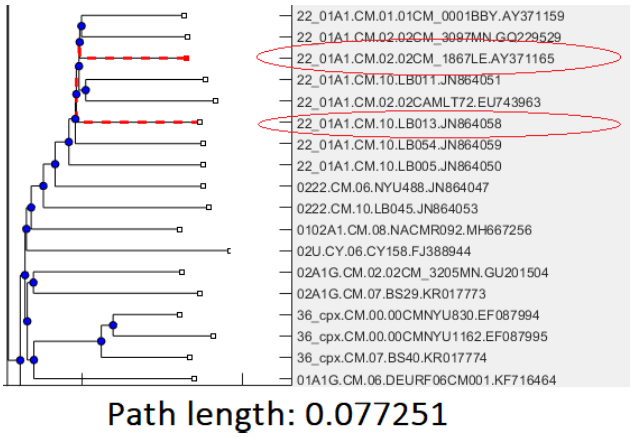
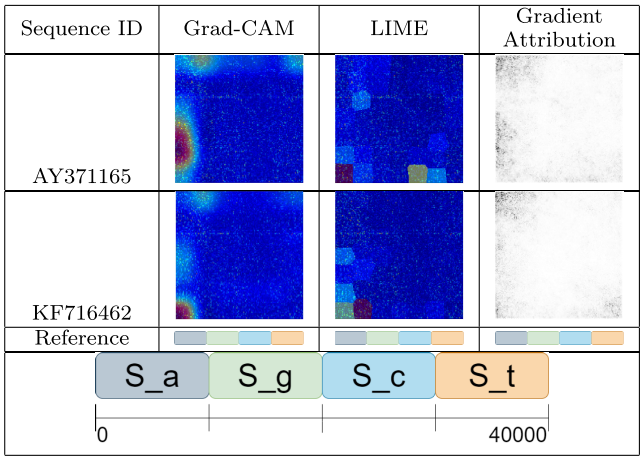


FIGURE 10. Phylogenetic Tree AY371165 and JN864058. These sequences are included in the recombinant subtype 22-01A1, which contains 21 sequences. As shown, both are phylogenetically close sequences, with a path distance of 0.077251. Their last common node results in a clade of seven sequences.

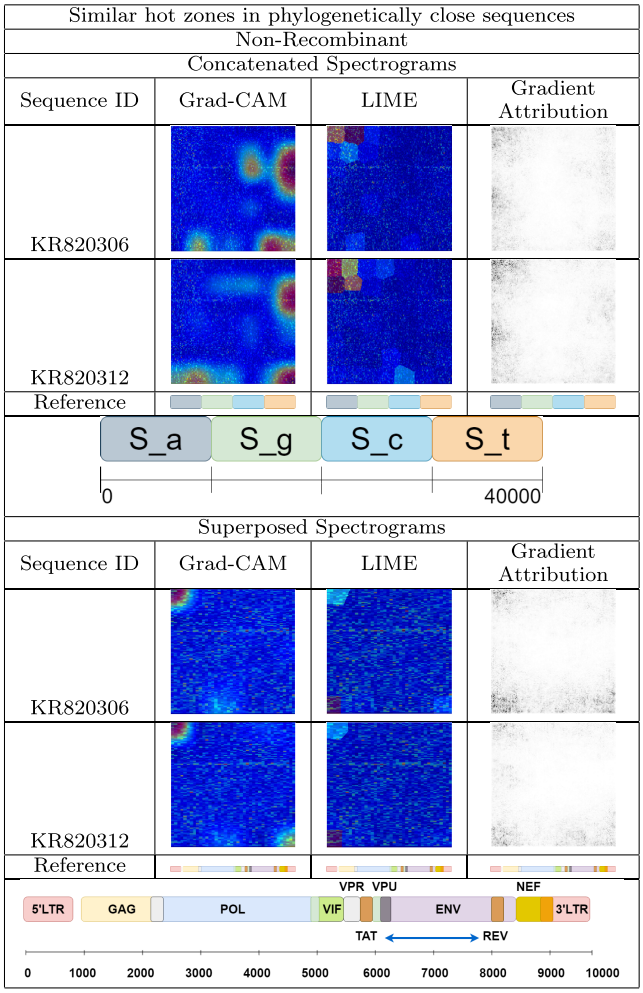
TABLE 6. Comparison of Interpretability Recombinant Concatenated Spectrograms - Subtype 22-01A1. AY371165 and KF716462 have close phylogenetic positions and their hot zones are similar. Grad-CAM points to low and one-sixth frequencies of A. Another hot zone is at high frequencies of A and the beginning of G. LIME basically points to low frequencies of A. Gradient Attribution is noisy but visually centered in A nucleotide area and T (low frequencies).



and KR820312, both belonging to subtype C and intra-patient (Patient ID ZM331F).

In some cases of recombinant spectrograms, hot areas are slightly more volatile, particularly in the concatenated graphical representation. In the case of AY371129 and AY371130, both from subtype 02-AG, their concatenated spectrograms (Table 8) show divergences between the interpretability hot zones despite being phylogenetically close sequences. Their path length is 0.085774. However, in the case of superposed spectrograms, their hot zones are similar. Given the phylogenetic closeness between both sequences, a higher level of similarity between hot zones in concatenated spectrograms is expected. Consistency occurs in the case of superposed spectrograms. This fact requires further research.

TABLE 7. Similar Hot Zones in Phylogenetically Close Non-Recombinant Sequences. KR820306 and KR820312 are intra-patient subtype C sequences. As expected, their decision areas are absolutely similar. In concatenated spectrograms, Grad-CAM highlights low and one-third frequencies of T nucleotide, with certain importance of one-third in C spectrogram. Another hot area is located at low frequencies of A and maybe G. LIME centers in high frequencies of A. Gradient Attribution shows more activity at A and T spectrograms. In superposed spectrograms, Grad-CAM fixes in high frequencies of 5'LTR in both cases. KR820312 has a secondary area at low frequencies of 3'LTR. A very peripheral area can be perceived at *pol* gene. LIME highlights low frequencies of 5'LTR, followed by high frequencies of this area. Gradient Attribution seems to emphasize high and low frequencies of 5'LTR and *gag*, and low frequencies of some area near *vif*, *env* and, specially, 3'LTR.



In general, there is a high similarity level of the hot zones as a function of phylogenetic closeness in concatenated and superposed spectrograms. In some cases, this match is even slightly higher in the superposed case.

An additional example is the case of intra-patient recombinant sequences MW443219 and MW443225 (Patient Code 40503), belonging to subtype 01-AE (Table 9).

As shown in Table 9, intra-patient (Patient ID 40503) MW443219 and MW443225 sequences (subtype 01-AE) develop the same high concordance in concatenated and superposed spectrograms.

Thus, based on the findings of the interpretability tools, the CNN exhibits consistent behavior for both datasets. This

TABLE 8. Comparison of Interpretability Recombinant Spectrograms - Subtype 02-AG. Grad-CAM hot zones are different in concatenated spectrograms. Hot zones in AY371129 are around high frequencies of A and one-ninth and high frequencies in T. However, AY371130 most relevant zones are one-sixth and high frequencies of A. LIME has no deviations. The most determinant areas are high and low frequencies in A. In Gradient Attribution, the most important region seems to be the high frequency range of A. In superposed, Grad-CAM relevant zone is located around one-third frequency of 5'LTR. Secondary areas are surrounding low frequencies of 5'LTR. LIME results are more dispersed. In AY371129, a zone close to one-ninth frequency at 5'LTR outlines. Less relevant sites are identified at low frequencies and one-third frequency at *gag*. In AY371130, the most determinant zone is one-third of 5'LTR. There is a secondary area in the low frequency range of the central part of the genome. Gradient Attribution point clouds are visually different, more noisy in the case of AY371129. In AY371129, the most relevant areas are low frequencies in 5'LTR and the first half of *gag*, *vif* and surroundings and 3'LTR. Another remarkable zone is between one-third and high frequencies of 5'LTR and *gag*. In AY371130, Gradient Attribution shows low activity, highlighting low frequencies of *gag* and maybe 5'LTR, around *vif* and, very weak, 3'LTR.

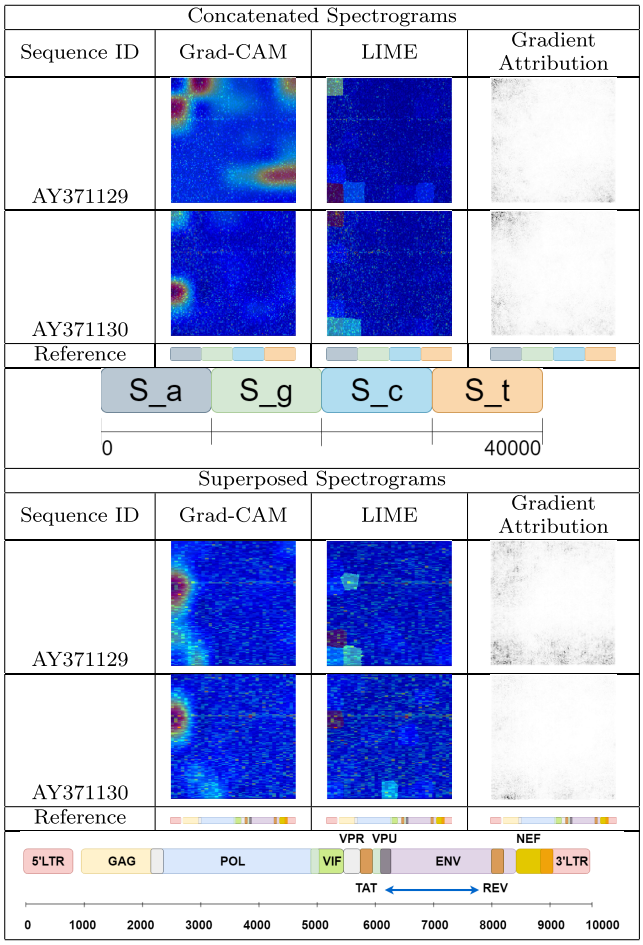
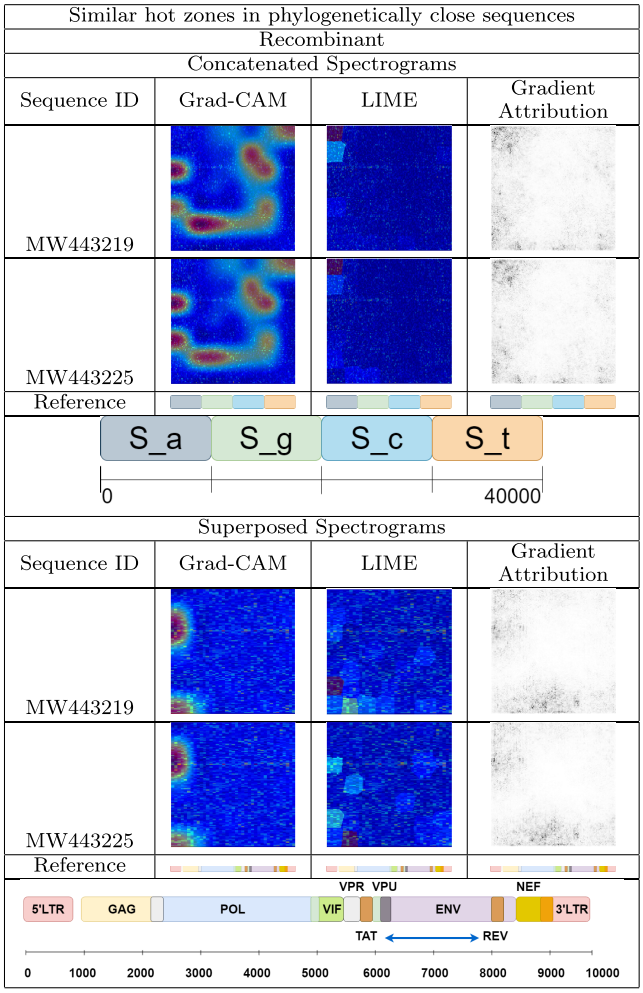


TABLE 9. Similar Hot Zones in Phylogenetically Close Recombinant Sequences. MW443219 and MW443225 are intra-patient recombinant sequences of subtype 01-AE. The most remarkable feature of Grad-CAM is the high similarity level of the hot zones in concatenated spectrograms. The main key areas are one-third, one-sixth and one-ninth frequencies in A, one-ninth frequency in G, some points near one-third, one-sixth and one-ninth in C, and one-sixth, one third and high frequencies of T. However, LIME fixes in high frequencies of A. Noisy results of Gradient Attribution seem to enhance low and high frequencies of A, some areas of low frequency range in G and C and low and one-third frequencies of T. In superposed spectrograms, Grad-CAM points to one-third and low frequencies of 5'LTR. LIME focuses at frequencies near one-ninth of 5'LTR and, low frequencies of *gag*. One minor zone is one-third in 5'LTR. Hot points in Gradient Attribution are grouped in one-third and high frequencies of 5'LTR, low frequencies in the central area of the genome and low frequencies around 3'LTR.



is particularly true for non-recombinant sequences, as hot zones tend to be more similar the closer the sequences are phylogenetically. In the case of recombinants there is greater volatility. Sometimes, a greater phylogenetic closeness does not correlate with a greater similarity in decision patterns. Compared to concatenated spectrograms, superposed spectrograms reflect the relationship between phylogenetic closeness and the level of similarity in interpretability tool results more accurately.

C. ANALYSIS OF FAILED CLASSIFICATIONS

A failed sequence is considered a sequence of the test dataset misclassified by the CNN. That is, a non-recombinant sequence is classified as recombinant and vice versa.

In several cases, the CNN fails the same sequence in both datasets. Specifically, non-recombinant FJ389363 and recombinant GU230127, HQ385479, JN864058, JX390977, MN172223 and MT559132 are double misclassified. These phenomena suggest that these sequences are biologically interesting, and have distinct features compared with the

TABLE 10. Failed Classification Analysis Example 1. FJ389363 (subtype G) is misclassified as recombinant. Comparing its hot areas with those of AJ508596 (recombinant sequence of subtype 30-0206) shows a high level of concordance between them, even in Gradient Attribution. Incorrect patterns are similar to those of the recombinant determinant areas. In concatenated spectrograms, Grad-CAM focuses on A spectrogram, LIME at low frequencies of A and Gradient Attribution at high frequencies of A and low frequencies of T and C. In superposed spectrograms, the Grad-CAM hot zones are low frequencies of 5'LTR. LIME is more dispersed throughout the low-frequency range, locating the main zones between 5'LTR and *gag* and at some points near *vif*. Gradient Attribution hot areas seem to focus on low and high frequencies.

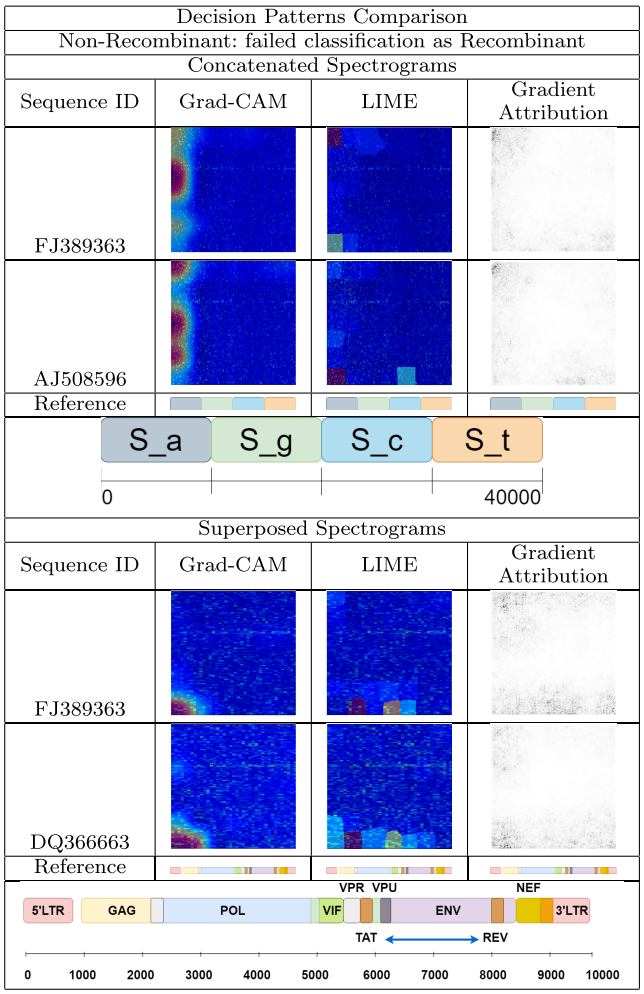
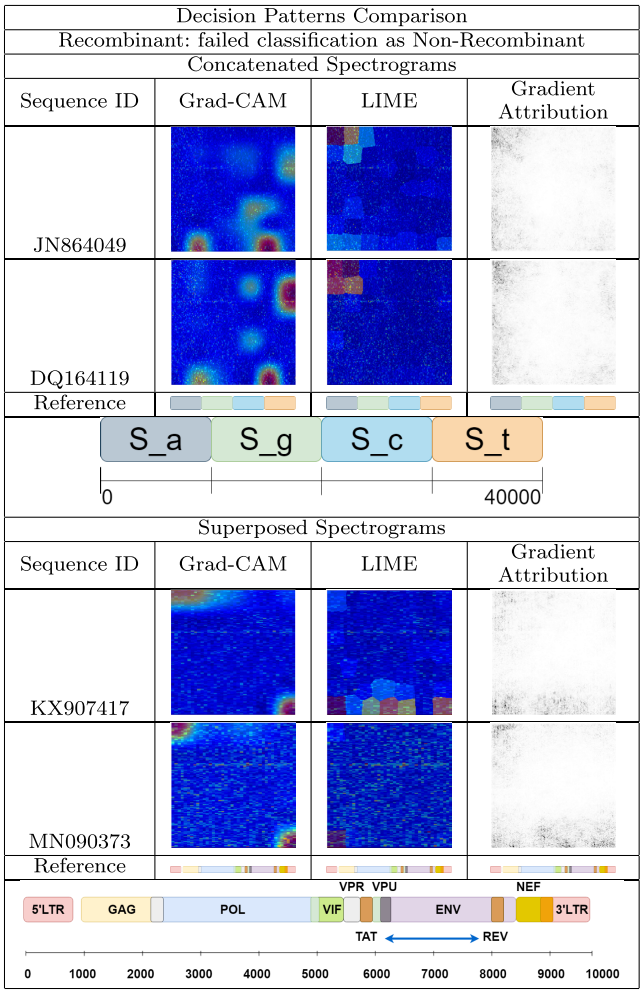


TABLE 11. Failed Classification Analysis Example 2. JN864049 (subtype 22-01A1) is misclassified as non-recombinant. Comparing its hot areas with those of DQ164119 (non-recombinant sequence of subtype C) shows again a high level of concordance between them, even in Gradient Attribution. In concatenated spectrograms, Grad-CAM highlights low frequencies between A and G and some zones near one-sixth, one-third and low frequencies in C and T. LIME focuses on high frequencies of A and Gradient Attribution is most active in the whole frequency range of A (specially high frequencies), followed by T (mostly one-third and low frequencies). In superposed spectrograms, Grad-CAM hot zones are high frequencies of 5'LTR and low frequencies of 3'LTR. LIME results are more scattered, but the determinant zone of LIME is low frequency range of 5'LTR. Additional areas are detected in low frequencies of the central part of the genome (end of *pol* and *vif*) and 3'LTR. Gradient Attribution determinant points are along the low frequency range and high frequencies of 5'LTR.



rest of the sequences. Therefore, a deep analysis is required regarding genome integrity and their correct classification in their subtypes. It is possible that these double misclassifications are not random, and that these sequences may have issues with their classification into their respective subtypes.

Analyzing non-recombinant sequences incorrectly classified as recombinant, the CNN seems to make a decision based on recombinant hot zones. Two representative examples are presented in Table 10.

The CNN incorrectly classifies the concatenated spectrogram of the non-recombinant sequence FJ389363 (subtype G) as a recombinant. Analyzing the Grad-CAM hot zones, we observe a high resemblance to AJ508596 hot zones (recombinant sequence of subtype 30-0206). To a lesser extent, the LIME hot points are also similar and Gradient Attribution is the most different, although they are still alike.

Similarly, the CNN also misclassifies the superposed spectrogram of the sequence FJ389363. In this case, the Grad-CAM and LIME hot zones in the two study sequences are very similar. In contrast, those of Gradient Attribution show appreciable differences to the naked eye. In any case, there is a direct link between the misclassification and hot zones of the incorrect category.

The same phenomenon occurs in the reverse direction. When the CNN fails to classify a recombinant sequence as non-recombinant, the decision zones look like those

associated with non-recombinants. Table 11 presents two graphical examples.

In Table 11, for concatenated spectrograms, recombinant JN864049 (subtype 22-01A1) is incorrectly classified as non-recombinant. Comparing the hot zones of its spectrograms with those of the non-recombinant sequence DQ164119 (subtype C), we observe that the decision patterns of the CNN when making the wrong decision are similar to those corresponding to DQ164119, both in concatenated and superposed. Even Gradient Attribution shows a higher degree of resemblance, which is greater than that seen in the previous example.

Similarly, the hot zones of the recombinant sequence KX907417 (subtype 41-CD), incorrectly classified as non-recombinant, are compared for the superposed spectrograms. High similarity is observed in the Grad-CAM hot zones corresponding to the non-recombinant MN090373 sequence (subtype B). In the case of LIME, the KX907417 hot zones are less specific than those in the case of MN090373. Similarities in the Gradient Attribution hot zones can be observed with the naked eye.

VI. CONCLUSION

Our methodology allows a rapid, robust and coherent identification of genomic areas that are determinant in the characterization of genetic recombination in HIV-1. This localization can significantly reduce genomic research timelines, and help microbiological experts to focus on the genomic areas. Our methodology allows for a complementary representation of genomic sequences, in addition to the traditional nucleotide sequence format. The conversion of genomic sequences to biologically consistent images such as spectrograms, enables the application of powerful deep learning and machine learning tools to the field of genomics, thus taking advantage of their full potential. The high hit rates on genomic spectrograms, despite their distinct nature from ImageNet images, demonstrate the transferability and applicability of the pattern detection capability from these tools to new domains. Therefore, we can efficiently identify areas of interest within the sequences through massive information comparison. This type of analysis is much more arduous to perform using the traditional system.

The CNN's total hit rate (test accuracy) for classifying viral sequences as recombinants or non-recombinants is 94% in both concatenated and superposed spectrograms. The decision areas in each category are consistent and differentiated. The hot zones are similar for sequences of the same subtype and phylogenetic proximity. When the CNN misclassifies, both Grad-CAM and LIME target the hot zones of the incorrect category.

The importance of the 5' and 3' LTR zones and their U3, R and U5 regions as hot zones for classifying a sequence as recombinant may be related to the role played by these terminal sections of the genome during viral transcription [82], [83], [84]. The neighboring *gag*, *nef* and *env* genes may also be involved in HIV-1 genetic recombination.

The importance of A and T nucleotides in recombinant feature correlates with the high content of polyA and polyT stretches in the HIV-1 sequence. These regions are susceptible to polymerase skipping and thus, genetic recombination. The key role of A nucleotide could be related to the high adenine content in lentivirus, especially in HIV-1 [85].

As Olabode et al. [86] suggested, the fact that sequences of different subtypes present very similar hot zones may be due to the existence of more cases of recombination than is currently described. It is possible that some viruses that are considered pure subtypes may exhibit hidden recombination phenomena.

In addition, the detected recombinant sequences are those which have prevailed because the resulting viruses have some selective advantages. Although numerous recombination phenomena can occur, they will not be selected if the resulting virus is non-viable or has lower fitness than the parental viruses. Through this research, we can gain insights into those genome regions that are more receptive to recombination and may confer a selective advantage for the virus [87].

Some sequences are misclassified in both concatenated and superposed spectrograms. The integrity and coherence of these published genomes must be verified. In addition, the criteria for classifying these sequences into their respective subtypes should be revised. This phenomenon may reinforce the correct operation of the CNN. If the subtypes of these sequences are incorrectly assessed or their genomes contain errors, it is biologically coherent that the CNN fails in both datasets.

In conclusion, 5'LTR and 3'LTR, and nucleotides A, followed by T, stand out as the leading hot zones for determining whether a sequence is recombinant or non-recombinant. These results seem biologically coherent with the genome structure and retrotranscription function. These facts validate that the CNN detects frequency-domain mathematical patterns that characterize a genomic sequence as recombinant or non-recombinant, including the location of these patterns within the sequence.

VII. FUTURE STUDIES

The application of methods based on multiple sequence alignment and manual adjustments did not allow for the discovery of patterns determining genetic recombination in viruses. Notwithstanding the information loss incurred through the conversion of sequences into spectrograms, it is possible to uncover previously unknown data embedded in the genome, thereby revealing patterns that are not visible in the nucleotide sequence itself.

The capacity of Deep Learning algorithms to find hidden patterns could greatly enhance research on this topic. To achieve this, it is necessary to delve into understanding the operation of these networks and how they detect these patterns. Detailed studies of existing pre-trained architectures and the development of interpretability tools are essential at this juncture. As more accurate neural networks are developed and applied to genomics, researchers will be able to

detect the specific regions of the genome that contribute to sequence characterization with enhanced precision and certainty. Consequently, the time required for sequence analysis and comparison will be significantly reduced. This holds particular significance for larger organisms with genomes spanning millions of nucleotides. Traditional analysis methods for such sequences are computationally intensive and extremely time-consuming (letter-by-letter manual adjustments). In such scenarios, it becomes crucial to possess a tool capable of pinpointing the genomic regions pertinent to the specific study feature. Our methodology can pave the way to address this need by offering precise identification, thereby greatly enhancing the efficiency of microbiological analyses.

Furthermore, it is necessary to feed these algorithms with curated datasets to help them identify these patterns. We needed to employ the entire dataset of available HIV-1 complete sequences for training and testing the neural network, which requires thousands of data samples to perform robustly. It is necessary to intensify the sequencing of diverse organisms to enhance the availability of the datasets.

This is the first time that a pre-trained CNN is applied to spectrogram datasets to detect and classify HIV-1 sequences as recombinant, with notable success rates and biologically supported results. Moreover, leveraging interpretability tools enabled us to identify specific genomic regions implicated in this feature. Our methodology contributes to the advancement of knowledge on genetic recombination in HIV-1, even though it may not have direct clinical application at present. Understanding the specific regions that play a key role in the selection of viable recombinant viruses can be useful for the search for new HIV-1 antiretroviral drugs. For example, the development of new antiretroviral treatments targeting specific genomic regions.

The UMI methodology is much faster and has lower computing costs than the traditional methods. This approach obviates the need for multiple sequence alignments and manual adjustments. These results indicate that there are probably mathematical patterns embedded in the genome that characterize a virus with different features. This encourages us to deepen their search and try to establish their mathematical formulations. Given its high accuracy, Gradient Attribution appears to be the most appropriate method for accurately locating these signatures.

We know where the CNN detects the recombinant feature. Our next step is to search for a signature or unique genetic identifier that allows to find similar variants or sequences in a much less computationally expensive manner.

Initiatives such as the one described, can successfully guide and reduce the heavy task of characterizing determining areas, which is probably applicable to more complex living organisms and their characteristics.

This research opens up new fields of application for biology.

Many exciting questions and fields emerge from this research.

VIII. DATA AVAILABILITY

All databases, codes and generated results are available from the corresponding author upon request. Supplementary Information is available at IEEE Dataport (Discovering Mathematical Patterns Behind HIV-1 Genetic Recombination: a new methodology to identify viral features - Supplementary Information).

ACKNOWLEDGMENT

The authors would like to thank a Professor Gerardo Mendizabal Ruiz with the Computer Science Department, University of Guadalajara, a Professor José Valentín Osuna Enciso with the Center for Exact Sciences and Engineering, University of Guadalajara, a Professor Ana Macarulla-Arenaza with the Faculty of Engineering, University of Deusto, a Research Associate Tony Castillo-Calzadilla with the DeustoTech Energy & Environment, a Research Assistant Maite Puerta-Beldarrain with the DeustoTech MoreLaboratory, a Research Assistant Armando Mendoza-Aguayo with DeustoTech Energy & Environment, and a Professor José Miguel Tamayo González with Loyola University Chicago for their support and assessment, also would like to thank the editors and reviewers for taking the time and effort necessary to review their manuscript, and also would like to thank all the valuable comments and suggestions, which helped them improve the quality of this article.

REFERENCES

[1] T. Hoenen et al., "Mutation rate and genotype variation of Ebola virus from Mali case sequences," *Science*, vol. 348, no. 6230, pp. 117–119, Apr. 2015.
[2] S. Duffy, "Why are RNA virus mutation rates so damn high?" *PLoS Biol.*, vol. 16, no. 8, Aug. 2018, Art. no. e3000003.
[3] J. Louten, "Virus replication," in *Essential Human Virology*, May 2016, ch. 4, pp. 49–70.
[4] M. M. C. Lai, "Genetic recombination in RNA viruses," in *Genetic Diversity of RNA Viruses*. Berlin, Germany: Springer, 1992, pp. 21–32.
[5] A. M. King, "Genetic recombination in positive strand RNA viruses," in *RNA Genetics*. Boca Raton, FL, USA: CRC Press, 2018, pp. 149–165.
[6] M. M. Lai, "RNA recombination in animal and plant viruses," *Microbiol. Rev.*, vol. 56, no. 1, pp. 61–79, Mar. 1992.
[7] C. Muslin, A. M. Kain, M. Bessaud, B. Blondel, and F. Delpeyroux, "Recombination in enteroviruses, a multi-step modular evolutionary process," *Viruses*, vol. 11, no. 9, p. 859, Sep. 2019.
[8] S. Su, G. Wong, W. Shi, J. Liu, A. C. K. Lai, J. Zhou, W. Liu, Y. Bi, and G. F. Gao, "Epidemiology, genetic recombination, and pathogenesis of coronaviruses," *Trends Microbiol.*, vol. 24, no. 6, pp. 490–502, Jun. 2016.
[9] A. D. Goldman and L. F. Landweber, "What is a genome?" *PLoS Genet.*, vol. 12, no. 7, 2016, Art. no. e1006181.
[10] R. G. Webster and W. Laver, "The origin of pandemic influenza," *Bull. World Health Org.*, vol. 47, no. 4, pp. 449–452, 1972.
[11] M. L. Kalish, K. E. Robbins, D. Pieniazek, A. Schaefer, N. Nzilambi, T. C. Quinn, M. E. S. Louis, A. S. Youngpairoj, J. Phillips, H. W. Jaffe, and T. M. Folks, "Recombinant viruses and early global HIV-1 epidemic," *Emerg. Infectious Diseases*, vol. 10, no. 7, pp. 1227–1234, Jul. 2004.
[12] P. Spreeuwenberg, M. Kroneman, and J. Paget, "Reassessing the global mortality burden of the 1918 influenza pandemic," *Amer. J. Epidemiol.*, vol. 187, no. 12, pp. 2561–2567, Dec. 2018.
[13] A. Mocroft, S. Vella, T. Benfield, A. Chiesi, V. Miller, P. Gargalianos, A. D. Monforte, I. Yust, J. Bruun, A. Phillips, and J. Lundgren, "Changing patterns of mortality across Europe in patients infected with HIV-1," *Lancet*, vol. 352, no. 9142, pp. 1725–1730, Nov. 1998.
[14] R. Nájera, E. Delgado, L. Pérez-Alvarez, and M. M. Thomson, "Genetic recombination and its role in the development of the HIV-1 pandemic," *AIDS*, vol. 16, pp. S3–S16, 2002.

[15] D. G. Higgins and P. M. Sharp, "CLUSTAL: A package for performing multiple sequence alignment on a microcomputer," *Gene*, vol. 73, no. 1, pp. 237–244, Dec. 1988.
[16] C. Notredame, D. G. Higgins, and J. Heringa, "T-Coffee: A novel method for fast and accurate multiple sequence alignment," *J. Mol. Biol.*, vol. 302, no. 1, pp. 205–217, Sep. 2000.
[17] C. B. Do, M. S. P. Mahabhashyam, M. Brudno, and S. Batzoglou, "ProbCons: Probabilistic consistency-based multiple sequence alignment," *Genome Res.*, vol. 15, no. 2, pp. 330–340, Feb. 2005.
[18] K. Katoh, G. Asimenos, and H. Toh, "Multiple alignment of DNA sequences with MAFFT," in *Bioinformatics for DNA Sequence Analysis*. Totowa, NJ, USA: Humana Press, 2009, pp. 39–64.
[19] R. C. Edgar, "MUSCLE: Multiple sequence alignment with high accuracy and high throughput," *Nucleic Acids Res.*, vol. 32, no. 5, pp. 1792–1797, Mar. 2004.
[20] Y. Liu, B. Schmidt, and D. L. Maskell, "MSAProbs: Multiple sequence alignment based on pair hidden Markov models and partition function posterior probabilities," *Bioinformatics*, vol. 26, no. 16, pp. 1958–1964, Aug. 2010.
[21] Y. Ye, D. W. Cheung, Y. Wang, S. M. Yiu, Q. Zhan, T. W. Lam, and H.-F. Ting, "GLProbs: Aligning multiple sequences adaptively," in *Proc. Int. Conf. Bioinf., Comput. Biol. Biomed. Informat.*, 2013, pp. 152–160.
[22] I. M. Wallace, O. Orla, and D. G. Higgins, "Evaluation of iterative alignment algorithms for multiple alignment," *Bioinformatics*, vol. 21, no. 8, pp. 1408–1414, Apr. 2005.
[23] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, pp. 436–444, May 2015.
[24] C.-C.-J. Kuo, "Understanding convolutional neural networks with a mathematical model," *J. Vis. Commun. Image Represent.*, vol. 41, pp. 406–413, Nov. 2016.
[25] H. Wang, A. Cruz-Roa, A. Basavanhally, H. Gilmore, N. Shih, M. Feldman, J. Tomaszewski, F. Gonzalez, and A. Madabhushi, "Mitosis detection in breast cancer pathology images by combining handcrafted and convolutional neural network features," *J. Med. Imag.*, vol. 1, no. 3, Oct. 2014, Art. no. 034003.
[26] H. Lu, Y. Zhu, M. Yin, G. Yin, and L. Xie, "Multimodal fusion convolutional neural network with cross-attention mechanism for internal defect detection of magnetic tile," *IEEE Access*, vol. 10, pp. 60876–60886, 2022.
[27] M. Rizwan, A. Shabbir, A. R. Javed, M. Shabbir, T. Baker, and D. A.-J. Obe, "Brain tumor and glioma grade classification using Gaussian convolutional neural network," *IEEE Access*, vol. 10, pp. 29731–29740, 2022.
[28] M. D. Kupperman, T. Leitner, and R. Ke, "A deep learning approach to real-time HIV outbreak detection using genetic data," *PLoS Comput. Biol.*, vol. 18, no. 10, Oct. 2022, Art. no. e1010598.
[29] S. Tammina, "Transfer learning using VGG-16 with deep convolutional neural network for classifying images," *Int. J. Sci. Res. Publ. (IJSRP)*, vol. 9, no. 10, pp. 143–150, 2019.
[30] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.
[31] S. Soffer, A. Ben-Cohen, O. Shimon, M. M. Amitai, H. Greenspan, and E. Klang, "Convolutional neural networks for radiologic images: A radiologist's guide," *Radiology*, vol. 290, no. 3, pp. 590–606, Mar. 2019.
[32] A. W.-C. Liew, H. Yan, and M. Yang, "Pattern recognition techniques for the emerging field of bioinformatics: A review," *Pattern Recognit.*, vol. 38, no. 11, pp. 2055–2073, Nov. 2005.
[33] A. E. Oueslati, N. Ellouze, and Z. Lachiri, "3D spectrum analysis of DNA sequence: Application to Caenorhabditis elegans genome," in *Proc. IEEE 7th Int. Symp. Bioinf. Bioeng.*, Oct. 2007, pp. 864–871.
[34] N. Dimitrova, Y. H. Cheung, and M. Zhang, "Analysis and visualization of DNA spectrograms: Open possibilities for the genome research," in *Proc. 14th ACM Int. Conf. Multimedia*, Oct. 2006, pp. 1017–1024.
[35] A. Bucur, J. van Leeuwen, N. Dimitrova, and C. Mittal, "Alignment method for spectrograms of DNA sequences," *IEEE Trans. Inf. Technol. Biomed.*, vol. 14, no. 1, pp. 3–9, Jan. 2010.
[36] S. Tiwari, S. Ramachandran, A. Bhattacharya, S. Bhattacharya, and R. Ramaswamy, "Prediction of probable genes by Fourier analysis of genomic sequences," *Bioinformatics*, vol. 13, no. 3, pp. 263–270, 1997.
[37] P. P. Vaidyanathan and B.-J. Yoon, "The role of signal-processing concepts in genomics and proteomics," *J. Franklin Inst.*, vol. 341, nos. 1–2, pp. 111–135, Jan. 2004.
[38] D. Sussillo, A. Kundaje, and D. Anastassiou, "Spectrogram analysis of genomes," *EURASIP J. Adv. Signal Process.*, vol. 2004, no. 1, Dec. 2004, Art. no. 790248.
[39] V. Kubicova and I. Provaznik, "Use of whole genome DNA spectrograms in bacterial classification," *Comput. Biol. Med.*, vol. 69, pp. 298–307, Feb. 2016.
[40] J. A. Morales, R. Saldaña, M. H. Santana-Castolo, C. E. Torres-Cerna, E. Borrayo, A. P. Mendizabal-Ruiz, H. A. Vélez-Pérez, and G. Mendizabal-Ruiz, "Deep learning for the classification of genomic signals," *Math. Problems Eng.*, vol. 2020, May 2020, Art. no. 7698590.
[41] H. J. Nussbaumer, "The fast Fourier transform," in *Fast Fourier Transform and Convolution Algorithms*. Cham, Switzerland: Springer, 1981, pp. 80–111.
[42] J. G. Proakis and D. G. Manolakis, *Introduction to Digital Signal Processing*. Englewood Cliffs, NJ, USA: Prentice-Hall, 1988.
[43] D. Ravì, C. Wong, F. Deligianni, M. Berthelot, J. Andreu-Perez, B. Lo, and G.-Z. Yang, "Deep learning for health informatics," *IEEE J. Biomed. Health Informat.*, vol. 21, no. 1, pp. 4–21, Jan. 2017.
[44] L. Yuan and J. Cao, "Patients' EEG data analysis via spectrogram image with a convolution neural network," in *Proc. Int. Conf. Intell. Decis. Technol.* Cham, Switzerland: Springer, 2017, pp. 13–21.
[45] K. N. Khan, F. A. Khan, A. Abid, T. Olmez, Z. Dokur, A. Khandakar, M. E. H. Chowdhury, and M. S. Khan, "Deep learning based classification of unsegmented phonocardiogram spectrograms leveraging transfer learning," *Physiol. Meas.*, vol. 42, no. 9, Sep. 2021, Art. no. 095003.
[46] G. Ruffini, D. Ibañez, M. Castellano, L. Dubreuil-Vall, A. Soria-Frisch, R. Postuma, J.-F. Gagnon, and J. Montplaisir, "Deep learning with EEG spectrograms in rapid eye movement behavior disorder," *Frontiers Neurol.*, vol. 10, p. 806, Jul. 2019.
[47] W. Zhang, J. Han, and S. Deng, "Heart sound classification based on scaled spectrogram and tensor decomposition," *Expert Syst. Appl.*, vol. 84, pp. 220–231, Oct. 2017.
[48] M. Qiao, Z. Fang, Y. Guo, S. Zhou, C. Chang, and Y. Wang, "Breast calcification detection based on multichannel radiofrequency signals via a unified deep learning framework," *Expert Syst. Appl.*, vol. 168, Apr. 2021, Art. no. 114218.
[49] M. Etemadi, S. B. Abkenar, A. Ahmadzadeh, M. H. Kashani, P. Asghari, M. Akbari, and E. Mahdipour, "A systematic review of healthcare recommender systems: Open issues, challenges, and techniques," *Expert Syst. Appl.*, vol. 213, Mar. 2023, Art. no. 118823.
[50] E. Santo and N. Dimitrova, "Improvement of spectral analysis as a genomic analysis tool," in *Proc. IEEE Int. Workshop Genomic Signal Process. Statist.*, Jun. 2007, pp. 1–4.
[51] M. Nahiduzzaman, M. R. Islam, and R. Hassan, "ChestX-ray6: Prediction of multiple diseases including COVID-19 from chest X-ray images using convolutional neural network," *Expert Syst. Appl.*, vol. 211, Jan. 2023, Art. no. 118576.
[52] P. M. Cheng and H. S. Malhi, "Transfer learning with convolutional neural networks for classification of abdominal ultrasound images," *J. Digit. Imag.*, vol. 30, no. 2, pp. 234–243, Apr. 2017.
[53] I. Kandel and M. Castelli, "Transfer learning with convolutional neural networks for diabetic retinopathy image classification. A review," *Appl. Sci.*, vol. 10, no. 6, p. 2021, Mar. 2020.
[54] A. Kensert, P. J. Harrison, and O. Spjuth, "Transfer learning with deep convolutional neural networks for classifying cellular morphological changes," *SLAS Discovery, Advancing Life Sci. R&D*, vol. 24, no. 4, pp. 466–475, 2019.
[55] H. W. Loh, C. P. Ooi, E. Aydemir, T. Tuncer, S. Dogan, and U. R. Acharya, "Decision support system for major depression detection using spectrogram and convolution neural network with EEG signals," *Expert Syst.*, vol. 39, no. 3, Mar. 2022, Art. no. e12773.
[56] D. Posada and K. A. Crandall, "Evaluation of methods for detecting recombination from DNA sequences: Computer simulations," *Proc. Nat. Acad. Sci. USA*, vol. 98, no. 24, pp. 13757–13762, Nov. 2001.
[57] Q.-S. Zhang and S.-C. Zhu, "Visual interpretability for deep learning: A survey," *Frontiers Inf. Technol. Electron. Eng.*, vol. 19, no. 1, pp. 27–39, Jan. 2018.
[58] S. Chakraborty, R. Tomsett, R. Raghavendra, D. Harborne, M. Alzantot, F. Cerutti, M. Srivastava, A. Preece, S. Julier, R. M. Rao, T. D. Kelley, D. Braines, M. Sensoy, C. J. Willis, and P. Gurram, "Interpretability of deep learning models: A survey of results," in *Proc. IEEE SmartWorld, Ubiquitous Intell. Comput., Adv. Trusted Comput., Scalable Comput. Commun., Cloud & Big Data Comput., Internet People Smart City Innov. (SmartWorld/SCALCOM/UIC/ATC/CBDcom/IOP/SCI)*, Aug. 2017, pp. 1–6.
[59] J.-X. Mi, A.-D. Li, and L.-F. Zhou, "Review study of interpretation methods for future interpretable machine learning," *IEEE Access*, vol. 8, pp. 191969–191985, 2020.

[60] V. Jahmunah, E. Y. K. Ng, R.-S. Tan, S. L. Oh, and U. R. Acharya, "Explainable detection of myocardial infarction using deep learning models with grad-CAM technique on ECG signals," *Comput. Biol. Med.*, vol. 146, Jul. 2022, Art. no. 105550.
[61] Y. Zhang, K. Song, Y. Sun, S. Tan, and M. Udell, "'Why should you trust my explanation?' Understanding uncertainty in LIME explanations," 2019, *arXiv:1904.12991*.
[62] K. Mridha, M. M. Uddin, J. Shin, S. Khadka, and M. F. Mridha, "An interpretable skin cancer classification using optimized convolutional neural network for a smart healthcare system," *IEEE Access*, vol. 11, pp. 41003–41018, 2023.
[63] I. E. Nielsen, D. Dera, G. Rasool, N. Bouaynaya, and R. P. Ramachandran, "Robust explainability: A tutorial on gradient-based attribution methods for deep neural networks," 2021, *arXiv:2107.11400*.
[64] Z. Salahuddin, H. C. Woodruff, A. Chatterjee, and P. Lambin, "Transparency of deep neural networks for medical image analysis: A review of interpretability methods," *Comput. Biol. Med.*, vol. 140, Jan. 2022, Art. no. 105111.
[65] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016.
[66] O. D. T. Catalá, I. S. Igual, F. J. Pérez-Benito, D. M. Escrivá, V. O. Castelló, R. Llobet, and J.-C. Peréz-Cortés, "Bias analysis on public X-ray image datasets of pneumonia and COVID-19 patients," *IEEE Access*, vol. 9, pp. 42370–42383, 2021.
[67] L. Chen, A. Cruz, S. Ramsey, C. J. Dickson, J. S. Duca, V. Hornak, D. R. Koes, and T. Kurtzman, "Hidden bias in the DUD-E dataset leads to misleading performance of deep learning in structure-based virtual screening," *PLoS ONE*, vol. 14, no. 8, Aug. 2019, Art. no. e0220113.
[68] V. Gupta and R. Patel, "Lungs disease classification using VGG-16 architecture with PCA," in *Proc. Int. Conf. Advancement Comput. Comput. Technol. (InCACCT)*, May 2023, pp. 495–500.
[69] Z. Zhou, Y. Liu, Q. Wang, and T. T. Toe, "Detection of pneumonia based on ResNet improved by attention mechanism," in *Proc. IEEE 3rd Int. Conf. Power, Electron. Comput. Appl. (ICPECA)*, Jan. 2023, pp. 859–863.
[70] R. Singh, N. Sharma, and R. Gupta, "Detection of Alzheimer's risk level using inception V3 transfer learning model," in *Proc. Int. Conf. Distrib. Comput. Electr. Circuits Electron. (ICDCECE)*, Apr. 2023, pp. 1–6.
[71] K. D. Hyde, D. Udayanga, D. S. Manamgoda, L. Tedersoo, E. Larsson, K. Abarenkov, Y. J. K. Bertrand, B. Oxelman, M. Hartmann, H. Kauserud, M. Ryberg, E. Kristiansson, and H. R. Nilsson, "Incorporating molecular data in fungal systematics: A guide for aspiring researchers," 2013, *arXiv:1302.3244*.
[72] E. Castro-Nallar, M. Pérez-Losada, G. F. Burton, and K. A. Crandall, "The evolution of HIV: Inferences using phylogenetics," *Mol. Phylogenetics Evol.*, vol. 62, no. 2, pp. 777–792, Feb. 2012.
[73] C. Kuiken, B. Korber, and R. W. Shafer, "HIV sequence databases," *AIDS Rev.*, vol. 5, no. 1, pp. 52–61, 2003.
[74] E. Simon-Loriere and E. C. Holmes, "Why do RNA viruses recombine?" *Nature Rev. Microbiol.*, vol. 9, no. 8, pp. 617–626, Aug. 2011.
[75] J. M. Coffin, S. H. Hughes, and H. E. Varmus, *Retroviruses*. Cold Spring Harbor, NY, USA: Cold Spring Harbor Laboratory Press, 1997.
[76] A. Santana, C. Domínguez, A. Lemes, T. Molero, and E. Salido, "Biología celular y molecular del virus de inmunodeficiencia humana (VIH)," *Revista Diagnóstico Biológico*, vol. 52, no. 1, pp. 7–18, 2003.
[77] R. Gómez-Román and C. Soler-Claudín, "La importancia de la secuencia terminal repetida larga (LTR) en la patogenia del virus de la inmunodeficiencia humana," *Revista Biomédica*, vol. 11, no. 1, pp. 61–71, 2000.
[78] W. S. Hu and H. M. Temin, "Genetic consequences of packaging two RNA genomes in one retroviral particle: Pseudodiploidy and high rate of genetic recombination," *Proc. Nat. Acad. Sci. USA*, vol. 87, no. 4, pp. 1556–1560, Feb. 1990.
[79] G. Turk, M. Carobene, A. Monczor, A. E. Rubio, M. Gómez-Carrillo, and H. Salomón, "Higher transactivation activity associated with LTR and Tat elements from HIV-1 BF intersubtype recombinant variants," *Retrovirology*, vol. 3, no. 1, pp. 1–12, Dec. 2006.
[80] J. Hemelaar, R. Elangovan, J. Yun, L. Dickson-Tetteh, I. Fleminger, S. Kirtley, B. Williams, E. Gouws-Williams, and P. D. Ghys, "Global and regional molecular epidemiology of HIV-1, 1990–2015: A systematic review, global survey, and trend analysis," *Lancet Infectious Diseases*, vol. 19, no. 2, pp. 143–155, 2019.
[81] R. A. Neher and T. Leitner, "Recombination rate and selection strength in HIV intra-patient evolution," *PLoS Comput. Biol.*, vol. 6, no. 1, Jan. 2010, Art. no. e1000660.
[82] W.-S. Hu and H. M. Temin, "Retroviral recombination and reverse transcription," *Science*, vol. 250, no. 4985, pp. 1227–1233, Nov. 1990.
[83] B. C. Ramirez, E. Simon-Loriere, R. Galetto, and M. Negroni, "Implications of recombination for HIV diversity," *Virus Res.*, vol. 134, nos. 1–2, pp. 64–73, Jun. 2008.
[84] E. R. de Arellano, J. Alcamí, M. López, V. Soriano, and Á. Holguín, "Drastic decrease of transcription activity due to hypermutated long terminal repeat (LTR) region in different HIV-1 subtypes and recombinants," *Antiviral Res.*, vol. 88, no. 2, pp. 152–159, Nov. 2010.
[85] J. P. Vartanian, A. Meyerhans, B. Asjö, and S. Wain-Hobson, "Selection, recombination, and G—A hypermutation of human immunodeficiency virus type 1 genomes," *J. Virology*, vol. 65, no. 4, pp. 1779–1788, Apr. 1991.
[86] A. S. Olabode, G. T. Ng, K. E. Wade, M. Salnikov, H. E. Grant, D. W. Dick, and A. F. Y. Poon, "Revisiting the recombinant history of HIV-1 group M with dynamic network community detection," *Proc. Nat. Acad. Sci. USA*, vol. 119, no. 19, May 2022, Art. no. e2108815119.
[87] L. R. Banner, J. G. Keck, and M. M. C. Lai, "A clustering of rna recombination sites adjacent to a hypervariable region of the peplomer gene of murine coronavirus," *Virology*, vol. 175, no. 2, pp. 548–555, Apr. 1990.

ANA GUERRERO-TAMAYO received the Bachelor of Engineering degree in electronics and industrial automation from the Faculty of Engineering, University of Deusto, in 2001, the master's degree in occupational risk prevention with a focus on occupational safety, industrial hygiene and ergonomics, and applied psychosociology, and the M.B.A. Executive degree from the Faculty of Economics and Business Administration, University of the Basque Country (UPV/EHU), in 2006. She is currently pursuing the Ph.D. degree in the application of deep learning algorithms in genomics with the Faculty of Engineering, University of Deusto.

She received a Special Mention for the Best Master's Thesis and Best Record of the 2005–2006 Promotion. She also completed a course in international business management from the School of Industrial Organization of Madrid (EOI), where she received the Special Mention to the Best Record of the Promotion. She held positions of responsibility in private companies related to the field of electrical MV/HV networks and in strategic consulting for multinational corporations. She is also a Research Assistant with DeustoTech, University of Deusto. Her research interests include applied mathematics in genomics, the analysis of gene expression and regulation, the prediction of protein structure and function, the analysis of massive sequencing data, and the development of new tools for genomic data analysis and visualization.

BORJA SANZ URQUIJO received the Ph.D. degree (cum laude) in information systems, in 2012, with a focus on malware detection in android mobile devices.

He is currently a Professor with the Faculty of Engineering, University of Deusto. He has been a Researcher with the DeustoTech, Computing Research Unit, since 2008, becoming a Head Researcher, from 2015 to 2018. He contributed to over 100 projects, including H2020 and national and private initiatives. Throughout these projects, he served as a researcher and a project manager in several of them. Some of his standout projects are Detecting and Analyzing Terrorist-Related Online Content and Financing Activities (DANTE), Security and Trust in the Information Society (CENIT SEGURA), and Models for Malware Propagation through Online Social Networks (MARSOL). He has published several book chapters and more than 60 articles in specialized national and international impact journals, such as *Logic Journal of IGPL*, *Electronic Commerce Research and Applications*, *Expert Systems with Applications*, and IEEE Access. His primary research interests include machine learning, big data, knowledge discovery, natural language processing, information retrieval, and the application of these techniques in the field of social sciences. He is also working on fairness, ethics, accountability, and the impact of artificial intelligence in society.

Dr. Urquijo won the Award for the Best Oral Presentation at the International Congress on Human Rights, Emerging Challenges 2018, with the paper: "Hic Sunt draconis: Human Rights and Big Data, analysis of an unexplored collaboration." He is also a member of the Observatory on the

Social and Ethical Impact of Artificial Intelligence (ODISEIA). For more information visit the link: (http://paginaspersonales.deusto.es/bosanz).

CONCEPCIÓN CASADO received the Graduate degree in biological sciences, in 1985.

She is currently researching HIV-1/AIDS pathogenesis. She has been with Instituto de Salud Carlos III (ISCIII), since 1987. Her early work was focused on diagnosing retroviral infections and evaluating commercial reagents for detection. In 1995, she joined the Molecular Virology Group, where she completed her Ph.D. thesis in biologic sciences on the molecular characterization of the Spanish HIV-1 epidemic and the intra-patient evolution of HIV-1. One of her most important results was the development of a method for dating viral nucleotide sequences based on the application of a relaxed molecular clock to viral evolution, which has since been widely used in the laboratory. This methodology applied to the viral evolution of HIV-1 in patients without clinical progression (LTNPs) led to the identification of a group of patients, the LTNP-elite controllers (LTNP-EC), which are characterized by the presence of ancestral viruses and lack of evolution. Since then, she has focused on studying the viral and host factors associated with LTNPs, including the accumulation of defective viral forms and non-progressive host factors. She was a part of a research team that demonstrated the relevance of viral environment functionality in clinical progression. Recently, this team described, for the first time, the functional cure in 3 LTNP-EC patients infected for over 30 years without progression. She contributed to the creation of LTNP and elite controllers (EC-RIS) cohorts of the RETICS AIDS NETWORK. She has published collaborative works characterizing LTNP-EC patients, who represent a model of functional cure for HIV-1 infection. Since 2016, she has been combining her research work with participation in the diagnosis and reference of retrovirus infections with the National Center of Microbiology (ISCIII). Since 2020, she has been responsible for the Molecular Virology Group. She has participated in numerous research projects (14). She has published numerous articles in international journals (56). She has participated in numerous national and international conferences (40).

MARÍA-DOLORES MORAGUES TOSANTOS received the Graduate degree in biological sciences from the University of the Basque Country (UPV/EHU), in 1979, and the Ph.D. degree, in 1989. She is currently a Full Professor and a Honorary Collaborator with the Faculty of Medicine and Nursing, UPV/EHU. She has been teaching activity in the degree in nursing (human physiology, microbiology, research methodology, and public health), since 1983. She obtained a predoctoral grant from the Basque Government (1980–1982), which included a six-month stay with the Department of Microbiology, California College of Medicine, University of California at Irvine, Irvine (Mechanism of inducible phenotypic drug resistance in Mucor racemosus). From 1980 to 1990, her work focused on the dimorphism of the fungus Aureobasidium pullulans. Since 1991, she has been studying the diagnosis and pathogenesis of invasive candidiasis, including the production of polyclonal and monoclonal antibodies and the characterization of virulence factors. In 2010, she became responsible for the Invasive Fungal Infection Study Group (GEIFI). Her current research interests include the diagnosis of fungal infections and resistance to antifungal compounds. She has collaborated in more than 30 research projects subsidized in public calls, with at least ten of them as a PI. In total, she has published nine books/chapters and 78 research articles in indexed journals, and she has signed more than 150 communications/presentations in national and international scientific meetings. She has been a member of several scientific societies, most notably the Spanish Society of Microbiology (SEM), since 1978, Spanish Association of Mycology (AEM), since 1997, International Society for Human and Animal Mycology (ISHAM), since 2008, American Society of Microbiology (ASM), since 2010, and Spanish Society of Infectious Diseases and Clinical Microbiology (SEIMC), since 2011. She was an Assistant Editor (2002–2014) and a member of the editorial board of the *Revista Iberoamericana de Micología* (ISSN: 1130-1406), since 2015, and an evaluator of manuscripts for several scientific journals. Between 2008 and 2017, she was a member of the Ethics Committee for Research Related to Human Beings (CEISH) of the Commission on Ethics in Research and Teaching (CEID), and the Director of the Department of Nursing, School of Medicine and Nursing, UPV/EHU, from 2019 to 2020.

ISABEL OLIVARES received the bachelor's degree in general medicine and surgery, and the Ph.D. degree from Universidad Autónoma de Madrid (UAM).

She has been a Research Scientist with Instituto de Salud Carlos III (ISCIII), since 1990. She is actively involved in virological research. First, she studied herpes simplex virus type 2 during her Ph.D. degree. As a Postdoctoral Fellow, she contributed to studies on the African swine fever virus. Since 1990, her research has been devoted to study several facets of the Human Immunodeficiency Virus Type I (HIV-1). Initially, she was mainly involved in molecular epidemiology, variability, and genetic characterization studies of HIV-1 isolates. She has participated in the ''WHO Network for HIV Isolation and Characterization'' as it was the first WHO Laboratory Network organized for HIV-1 studies, with great relevance. Since 1996, she has been focusing on HIV-1 infection pathogenesis by ''in vitro'' experiments. She has generated and characterized two infectious molecular clones from a patient HIV-1 isolate. This is a remarkably useful tool for multiple ''in vitro'' studies. She has collaborated with Centro de Biología Molecular Severo Ochoa on three coordinated research projects to study the implications of the viral reverse transcriptase domain for reverse transcriptase copy fidelity, viral replicative capacity, and viral resistance to antiretrovirals. Since 2000, she has also been focused on studying HIV-1 persistence using various persistently infected cell lines obtained in the laboratory as a model to identify viral or cellular factors associated with viral persistence. She also explored lethal mutagenesis, a new antiviral strategy to extinguish HIV-1 through elevated mutation rate in a persistently infected cell line, in two coordinated research projects. Since 2003, she has been a member of the RETICS AIDS NETWORK, focusing on the viral and host features of LTNP patients. Recently, she has been interested in the search for new treatments against HIV-1 targeting host cell proteins, studying the antiviral effect of an Ataxia telangiectasia, and Rad3-related protein (ATR) inhibitor on HIV-1 replication. She has participated in numerous research projects and she is the author of 42 peer-reviewed articles.

IKER PASTOR-LÓPEZ received the bachelor's degree in computer engineering, in 2007, the master's degree in information security, in 2010, and the Ph.D. degree (Hons.) in computer science from the Faculty of Engineering, University of Deusto, in 2013.

He is currently an Assistant Professor and an Erasmus Coordinator with the Faculty of Engineering, University of Deusto, teaching courses, such as deep learning, computer vision, and information security, where he is also a Researcher with DeustoTech. He is a member of the Department of Information Technology, Electronics, and Communication. He is also the Director of the Master's Program in Computing and Intelligent Systems, Faculty of Engineering, University of Deusto, where he has been a member of the Big Data and Business Intelligence Program, since 2016. His research interests include big data analytics, deep learning, natural language processing, opinion mining, and computer vision. His published work includes more than 45 scientific articles and international conferences. He has participated in the conceptualization and scientific and technical development of numerous competitive projects and contracts with companies, the latter with several successful cases of knowledge transfer actions.

Dr. Pastor-López is a member of the Scientific Committee of several congresses, such as CISIS, SOCO, and ICEUTE. He is a Reviewer for many journals, including those indexed in JCR, such as *Neurocomputing*, *Logic Journal of IGPL*, and the *Engineering and Industry* magazine (DYNA), among others.