PLOS COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

In silico prediction of HIV-1-host molecular interactions and their directionality

Haiting Chai, Quan Gu, Joseph Hughes, David L. Robertson*

MRC-University of Glasgow Centre for Virus Research, Glasgow, United Kingdom

* david.l.robertson@glasgow.ac.uk

Abstract

Human immunodeficiency virus type 1 (HIV-1) continues to be a major cause of disease and premature death. As with all viruses, HIV-1 exploits a host cell to replicate. Improving our understanding of the molecular interactions between virus and human host proteins is crucial for a mechanistic understanding of virus biology, infection and host antiviral activities. This knowledge will potentially permit the identification of host molecules for targeting by drugs with antiviral properties. Here, we propose a data-driven approach for the analysis and prediction of the HIV-1 interacting proteins (VIPs) with a focus on the directionality of the interaction: host-dependency versus antiviral factors. Using support vector machine learning models and features encompassing genetic, proteomic and network properties, our results reveal some significant differences between the VIPs and non-HIV-1 interacting human proteins (non-VIPs). As assessed by comparison with the HIV-1 infection pathway data in the Reactome database (sensitivity > 90%, threshold = 0.5), we demonstrate these models have good generalization properties. We find that the 'direction' of the HIV-1-host molecular interactions is also predictable due to different characteristics of 'forward'/pro-viral versus 'backward'/pro-host proteins. Additionally, we infer the previously unknown direction of the interactions between HIV-1 and 1351 human host proteins. A web server for performing predictions is available at http://hivpre.cvr.gla.ac.uk/.

Author summary

Human immunodeficiency virus type 1 (HIV-1) is the cause of acquired immunodeficiency syndrome (AIDS), a disease with no effective cure despite decades of research. A better understanding of the molecular interactions between HIV-1 and human host proteins can facilitate the discovery of potential host targets which may be of great importance for the development of antiviral drugs that go beyond mere control of infection. In this study, we elucidate some host-dependency and antiviral factors that may be helpful to distinguish HIV-1 interacting human proteins (VIPs) from non-HIV-1 interacting human proteins (non-VIPs). We also consider the 'directionality' in the HIV-1-host protein-protein interactions, i.e., whether the interaction with the host molecule is in the interest of the virus or part of the anti-viral response. We design a machine learning framework to generate models based on the known information and use them for the

OPEN ACCESS

Citation: Chai H, Gu Q, Hughes J, Robertson DL (2022) In silico prediction of HIV-1-host molecular interactions and their directionality. PLoS Comput Biol 18(2): e1009720. https://doi.org/10.1371/journal.pcbi.1009720

Editor: Joel O. Wertheim, University of California San Diego, UNITED STATES

Received: May 31, 2021

Accepted: December 3, 2021

Published: February 8, 2022

Data Availability Statement: Source code is available at https://github.com/HChai01/HIVPRE. Data and an implementation are available at http://hivpre.cvr.gla.ac.uk/.

Funding: HC received finding from the China Scholarship Council under Grant 201706620069. JH, QG and DLR are funded by the Medical Research Council (MC_UU_1201412). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

classification of VIPs and non-VIPs. Our predictions have the potential to provide refined sets of human host targets aiding in the discovery of novel HIV-1 therapeutics.

Introduction

Human immunodeficiency virus type 1 (HIV-1) is the cause of acquired immunodeficiency syndrome (AIDS) and constitutes a major cause of human disease and associated comorbidities. Virus infection involves viral molecules exploiting the host cell in order to replicate. The engagement of the HIV-1 envelope glycoprotein and cell-surface receptors, CD4 and either the membrane-spanning C-C motif chemokine receptor 5 (CCR5) or C-X-C motif chemokine receptor 4 (CXCR4), initiates virus attachment and entry into the cell [1–3]. Virus molecules including the HIV-1 regulatory factors (*tat* and *rev*) and accessory proteins (*vpr*, *vif*, *nef*, and *vpu*) ensures viral persistence, replication, dissemination and transmission by modulating the surface and intracellular environment of the infected cell [4–8]. The production of HIV-1 *gag*/*pol* polyproteins is essential for assembly, release and maturation of new virions [9]. Protein-protein interactions (PPIs) between virus and host molecules enable the virus to infect and exploit host cell sub-systems to replicate and persist despite the host immune response [1–3,6–11]. Conversely, there are many human host molecules that function as antiviral factors and are part of the immune response [12–14]. Improving our understanding of these HIV-1-host PPIs can provide insights into the molecular mechanisms underlying virus persistence and pathogenesis. Determining the nature of virus-host interactions [15] is thus of importance for the discovery of potential host inhibitors or targets for antiviral therapeutics [16] exemplified by the CCR5 antagonist maraviroc [17]. Intuitively, there are many more possible drug-targets in the host compared to HIV's compact genome, which codes for relatively few proteins. To efficiently direct laboratory experiments and make use of rapidly accumulating data in the post-genomic era, the development of efficient *in silico* approaches has become an important area of research focus.

Over the past few years, several computational studies on HIV-1 have characterised attributes of HIV-1 interacting human proteins based on various data, e.g., gene ontology (GO) annotations [18], interaction network profiles [19], disease pathways [20] and post-transcriptional modification profiles [21]. A hierarchical biclustering system has been used [15] to designate HIV-1-host PPIs directionality, polarity and control properties. This research demonstrates how the HIV-1 interacting human proteins (VIPs) can be grouped by related virus-associated perturbations and can be distinguished from the non-HIV-1 interacting ones (non-VIPs). Curation of the extensive experimental literature has permitted an HIV-1-host PPI dataset to be compiled [18]. This can be used for the purpose of modelling and predictions via machine learning algorithms. For example, a random forest (RF) model was constructed by including 35 features for the prediction of HIV-1-host interaction pairs [22]. Further work integrated semi-supervised learning, multi-task learning and neural networks [23]. Subsequently, a biclustering-based approach was applied along with an association rule mining technique [24,25]. Supervised machine learning methods [26,27] have also been implemented using the support vector machine (SVM) algorithm and datasets with different positive-to-negative ratios. Based on the assumption that proteins with similar sequence or structural properties tend to share common interaction partners, studies have also predicted possible HIV-1-host interaction pairs by integrating protein short linear motifs (SLiMs) or protein structure information [28–30].

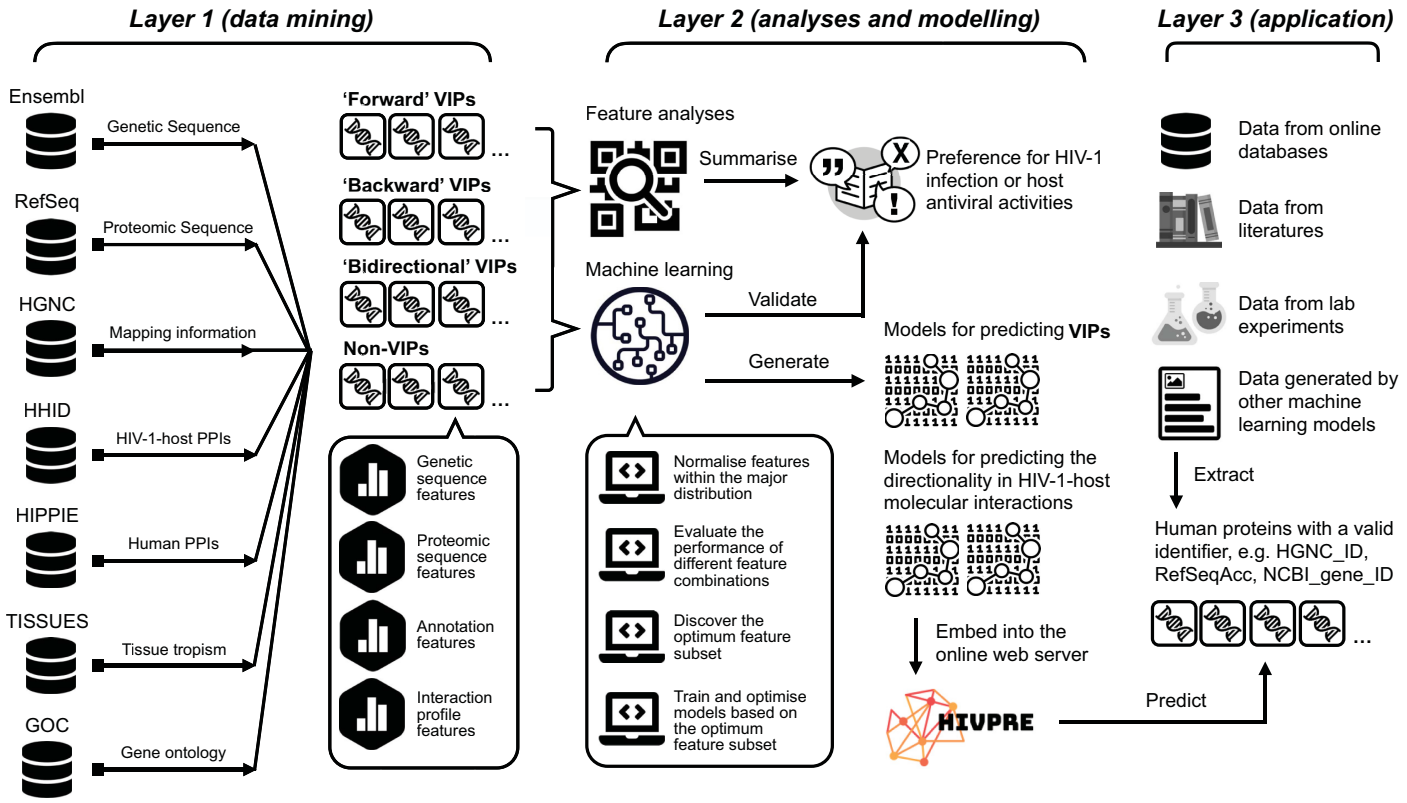

Fig 1. Diagrammatic representation of the project pipeline separated into three procedural layers. The figure is created using images from Wikimedia Commons, https://commons.wikimedia.org. Abbreviations: HIV-1, human immunodeficiency virus type 1; VIPs, HIV-1 interacting human proteins; non-VIPs, non-HIV-1 interacting human proteins; HGNC, HUGO Gene Nomenclature Committee; HHID, the HIV-1 Human Interaction Database; HIPPIE, Human Integrated Protein-Protein Interaction rEference; GOC, Gene Ontology Consortium.

https://doi.org/10.1371/journal.pcbi.1009720.g001

Owing to the contribution made by computational approaches [31–34], it is possible to obtain a list of potential VIPs with high confidence. However, there are still many improvements that can be made. The majority of the published methods [22–30] are highly dependent on the use of limited types of properties of the interacting HIV-1-host molecules. Some of the defined non-VIPs could be false negatives relative to different HIV-1 proteins [35]. For example, non-*env*-interacting protein cyclin T1 (CCNT1) interacts with HIV-1 during infection as it is targeted by *gag* and *tat* proteins [4,36,37]. The specific nature of the molecular interaction is important for understanding pro-viral interactions versus host antiviral activities. Crudely this can be broken down to the 'directionality' of the interaction [15]: 'forward'/pro-viral versus 'backward'/pro-host proteins, a prediction task addressed for the first time in this study. Additionally, there also exists a group of human host proteins having both pro-viral and pro-host properties, i.e., are 'bidirectional' in nature, for example, CD4 [38,39]. Finally, although the expansion of feature coverage provides a clearer picture for the classification problem, it also induces a series of problems such as feature redundancy [40] and overfitting [41–43].

To address these points, we propose a computational approach for the analysis and prediction of HIV-1-host molecular interactions (presented diagrammatically in **Fig 1**). Contrary to previous prediction-based studies [22–30], we introduce a broader definition for the HIV-1 interacting proteins. Human proteins targeting or being targeted by one or multiple HIV-1 proteins are all referred to as VIPs. Non-VIPs represent those human proteins without any

record of being directly involved in an HIV-1 interaction. We designed three procedures to maximise the set of the non-VIPs and to reduce their chance of being false negatives. Three tags: 'forward' (pro-viral), 'backward' (pro-host) and 'bidirectional' (pro-viral and pro-host) were assigned to VIPs to capture the direction of the virus-host interaction during the HIV-1 life cycle [15,44]. In total, we encoded 671 features based on the data retrieved from multiple databases [44–50] to characterise the human proteins by genetic, transcriptomic, proteomic and network information. We also measured the contribution of individual features and different feature combinations. We constructed different feature sets via two feature selection schemes to generate prediction models on the training datasets with the SVM method [51]. Performance on the testing datasets demonstrates good prediction quality and generalization capability of our VIP prediction models. A web server for HIV-1-host molecule prediction is available at http://hivpre.cvr.gla.ac.uk/.

Methods

Dataset curation

We retrieved 16215 HIV-1-host PPI records from the HIV-1 Human Interaction Database (HHID) (https://www.ncbi.nlm.nih.gov/genome/viruses/retroviruses/hiv-1/interactions/) [44] involving 7120 HIV-1-host interaction pairs and 3854 distinct VIPs (**S1 Data**). Protein sequences for the VIPs were collected from the NCBI's RefSeq database [46]. To avoid over-representation of similar protein sequences in the dataset, we grouped them into 2881 clusters using CD-HIT [52,53] with a threshold of 40% sequence similarity [52,53], and picked the longest sequence in each cluster as representative. This was to prevent producing feature vectors with high similarity, biasing the prediction performance. These 2881 representative VIPs formed our positives in dataset S1 (**Table 1**).

Table 1. Breakdown of VIP and non-VIP datasets used.

Dataset[a]		Positives	Negatives
Main dataset S1		2881 VIPs	7261 non-VIPs
	Training S1'	2304 VIPs	2304 non-VIPs
	Independent testing S1"	577 VIPs	4957 non-VIPs
Main dataset S2		188 backward VIPs	1007 forward VIPs
	Training S2'	150 backward VIPs	150 forward VIPs
	Independent testing S2"	38 backward VIPs	857 forward VIPs
Reference dataset S3		335 bidirectional VIPs	
Blind testing dataset S4		1351 undefined VIPs	
Testing dataset S5		234 VIPs	
Testing dataset S6		356 VIPs	

[a]Dataset S1 and S2 were constructed for the prediction of VIPs and their directionality in the HIV-1-host PPIs. 80% of positives and an equal number of negatives were randomly selected for training while the remaining 20% of proteins were used for testing. Dataset S3 was constructed for prediction of 'bidirectional' VIPs while S4 was constructed for the prediction of putative forward, backward or bidirectional VIPs. Testing datasets S5 and S6 were retrieved from two resources with high experimental confidence: the HIV-1 infection pathway in Reactome [60], https://reactome.org/PathwayBrowser/#/R-HSA-162906 and viral host-dependency epistasis map linked to the HIV function [61]. The lists of proteins sampled for training and independent testing are provided in **S2 Data**. Abbreviations: HIV-1, human immunodeficiency virus type 1; VIPs, HIV-1 interacting human proteins; non-VIPs, non-HIV-1 interacting human proteins.

https://doi.org/10.1371/journal.pcbi.1009720.t001

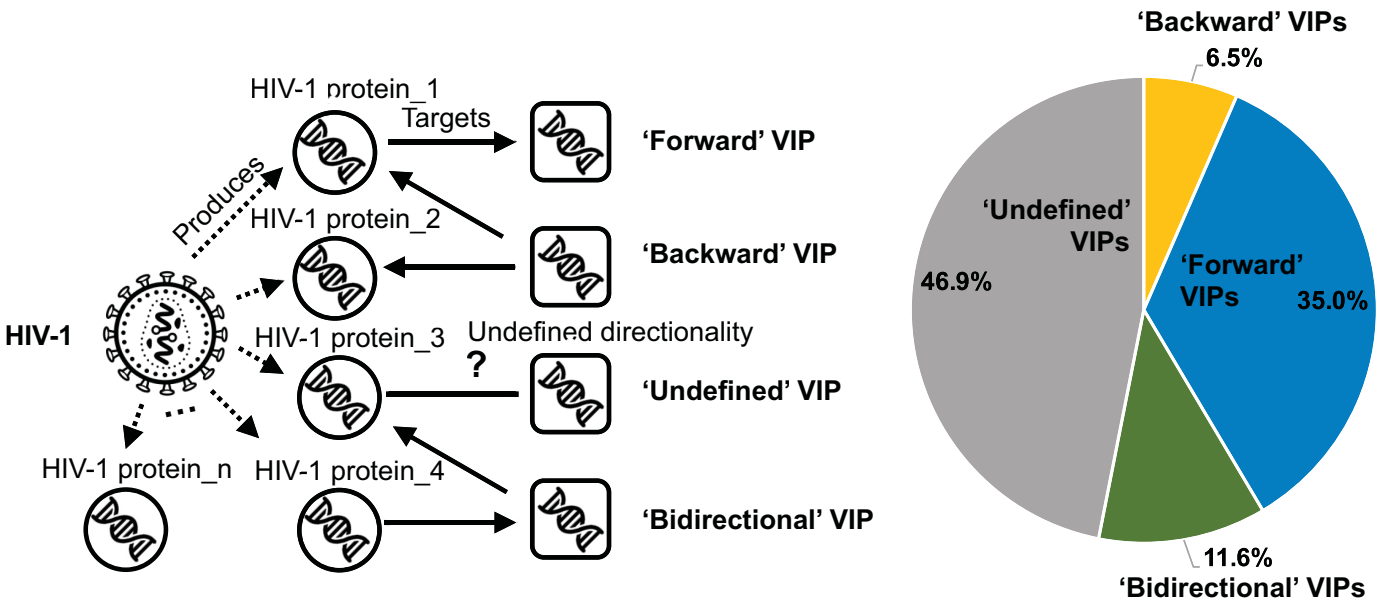

Fig 2. Representation of the characterisation of the types of virus interacting proteins (VIPs). VIPs were tagged as 'forward', 'backward' or 'bidirectional' based on the key words describing their interaction with HIV-1 proteins [44] and directionality designated by MacPherson *et al.* (https://doi.org/10.1371/journal.pcbi.1000863.s004) [15]. The direction was classed as 'undefined' if this information is not available. The direction tag for each VIP is provided in **S1 Data**. The figure is created using BioRender, https://biorender.com/. Abbreviations: HIV-1, human immunodeficiency virus type 1; VIP, HIV-1 interacting human protein; PPI, protein-protein interactions.

https://doi.org/10.1371/journal.pcbi.1009720.g002

Following the methods of MacPherson *et al.* [15], we assigned the VIPs direction tags: 'forward', 'backward' or both/'bidirectional' (**Fig 2**). The forward VIPs (e.g., C-X-C motif chemokine ligand 10, CXCL10) are pro-viral proteins. These host molecules are targeted by HIV-1, so-called host-dependency factors [54] and have no recorded antiviral response to the infection during the virus life cycle [55]. The backward VIPs (e.g., apolipoprotein L1, APOL1) are pro-host proteins that are associated with control or inhibition of the viral infection [56]. The bidirectional VIPs (e.g., CXCR4) are targeted by HIV-1 (forward direction) and can produce pro-host responses (backward direction) by influencing the same or different HIV-1 proteins during the viral infection. Some of these are potential therapeutic targets to inhibit virus replication by making host molecules unavailable to the virus [57,58]. Since the 'direction' of some HIV-1-host molecular interactions have not been clearly defined, these VIPs were not included in our analysis. Collectively, we obtained 188 (~6.5%) backward, 1007 (~35.0%) forward, 335 bidirectional (~11.6%) and 1351 (~46.9%) undefined VIPs from dataset S1 to construct another three datasets, S2, S3 and S4 for direction-related predictions (**Table 1**).

We performed three procedures to improve the quality of the non-VIP dataset and reduce potential false negatives. First, we chose human proteins produced by the canonical transcript since these proteins are assumed to express the main function of the gene [59]. Second, human proteins sharing more than 40% sequence similarity with any of the reported 3854 VIPs in the HHID were excluded to prevent sequences similar to the known VIPs overly influencing the modelling and predictions. Third, we controlled for the sequence similarity of non-VIPs at a 40% level as we did for VIPs. This is intended to prevent predictions being influenced by similar combinations of feature vectors. As a result, we obtained 7261 nonredundant non-VIPs to form the negatives in dataset S1 (**Table 1**).

As we applied different criteria compared to others to classify our VIPs/non-VIP datasets, it was hard to make direct comparisons between our predictions and previous studies [22–30]. We introduced two testing datasets consisting of VIPs with high experimental confidence from Reactome [60] and Gordon *et al*. [61] in order to assess the generalization capability of our machine learning models. A breakdown of the VIPs and non-VIPs used in this study is listed in **Table 1** and more detailed information is provided in **S2 Data**.

Feature generation

In this study, we encoded 671 different features mainly from six online databases: Ensembl [47], RefSeq [46], TISSUES [48], Human Integrated Protein-Protein Interaction rEference (HIPPIE) [50], HHID [44] and the Gene Ontology Consortium (GOC) [49]. Among them, 537 features were used to distinguish the VIPs from non-VIPs while 584 features were used to investigate the directionality of the HIV-1-host molecular interactions. Based on the data sources, our encoded features could be divided into four groups: (1) genome-based sequence, (2) proteome-based sequence, (3) annotation-based and (4) interaction profile-based features. The source code for generating these features is available at: https://github.com/HChai01/HIVPRE.

Genome-based sequence features. We compiled 107 genome-based sequence features for each human protein which included alternative splicing, nucleotide composition, codon usage and a measure of evolutionary conservation. Information in the alternative splicing data was encoded into four features to represent the evolution of phenotypic complexity in human genes [62,63]: the number of transcripts, protein-coding transcripts, exons and unit exon in transcripts (UET). Nucleotide composition represented the distribution of four basic nucleobases and their phosphodiester bonds-combinations, e.g., CpG, in the coding region of genetic sequences [64]. The usage of the existing 64 codons was calculated in each nucleotide sequence to reflect the balance between mutational biases and natural selection for translational optimization in different classes [65]. For evolutionary conservation, we collected the data from BioMart [47] and calculated the number of paralogues, synonymous substitutions (ds), non-synonymous substitutions (dN) and the ratio of dN to dS within human paralogues and orthologues in four homininae genomes: chimpanzee, gorilla, orangutan and gibbon. These features were used to assess the evolutionary selection pressure acting on the protein sequences [66].

Proteome-based sequence features. We encoded 251 features from proteome-based sequence data for the prediction of VIPs and their directionality. Discrete sequence information was calculated as amino acid compositions, while linear information was analysed from the perspective of SLiMs and intrinsic disorder. We generated 37 types of amino acid composition based on the differences in individual amino acids or their physiochemical attributes [67]. Ambiguous or other types of amino acids, e.g., selenocysteine, pyrrolysine etc. were masked as 'X' and ignored in this study. We used MERCI [68] to detect conserved sequence patterns as a result of strong purifying selection [69], obtaining 206 motifs representing putative SLiMs overrepresented in the group of VIPs and backward VIPs (Pearson's Chi-squared test, $P<0.05$). The occurrence of these potential SLiMs was then split and encoded into 206 non-parametric features with a binary system. Four features measuring the overall representation of VIP- or backward VIP-enriched SLiMs were added as hedges against random error caused by data imbalance [70]. The disordered regions in human protein sequences were identified using Espritz [71] and IUPred [72] as such regions have been linked to VIPs [73].

Annotation-based features. We encoded 292 annotation-based features with a binary system from the collected tissue and gene expression data. Among these, 66 features were generated by mapping the GO terms to the child term of three main GO root terms: molecular

function (GO:0003674), cellular component (GO:0005575) and biological process (GO:0008150) [49]. They characterise the domain in which human proteins may be involved such as binding (child term of molecular function, GO:0005488), intracellular (child term of cellular component, GO:0005622) and metabolic process (child term of biological process, GO:0008152) when interacting or not interacting with HIV-1 molecules [18]. The remaining 226 features were encoded based on the experimentally verified expression data in TISSUES [48] to reflect the association between tissue tropism and HIV-1 infection at a molecular level [74].

Interaction profile-based features. Interaction profile-based features were generated from HIV-1-host PPIs [44] and the human interactome [50]. We used 11 features to represent the degree to which a known VIP was central to the life cycle of HIV-1 [1–3]. Specifically, one feature was encoded to count the number of HIV-1 gene-products interacting with human host molecules and the remaining ten were binary-encoded to capture the interaction relationship between the host molecule and the corresponding HIV-1 gene-product, e.g., *gag*, *tat* or the antisense protein gene *asp* [75]. We retrieved 332,701 experimentally verified human-human PPIs with confidence scores higher than 0.63 involving 17,607 human proteins from HIPPIE [50] to pinpoint proteins with potential pathological or therapeutic relevance [76,77]. NetworkAnalyzer [78] was used to calculate ten different network features including the average shortest distance, degree, neighbourhood connectivity, betweenness, stress, closeness, eccentricity, radiality, topological coefficient and clustering coefficient. Human proteins not involved in the human-human PPI network were assigned zero values for all of the aforementioned network features.

Supervised machine learning and feature selection

We applied a supervised machine learning method for the prediction tasks. We used the SVM model with the radial basis function [51] after comparing it with the k-nearest neighbors (KNN), decision tree (DT) and random forest (RF) algorithms [33]. The SVM algorithm aims to find an appropriate hyperplane in the feature space for classifying the majority of positive and negative samples. It can tolerate the existence of some noisy or incorrect data but may be biased by different feature scales or imbalanced positive-to-negative ratios as it was designed to calculate the margin of the data [79]. Additionally, although the SVM algorithm can map the current feature space to a higher dimensional one for better classification [51], it is a sub-optimal strategy for including too many features for modelling even if they are all instructive. This can result in overfitting of the machine learning model [42] leading to a loss of robustness [80]. To address these points, we first used an undersampling strategy [70] to randomly construct balanced training datasets (**S2 Data**). Second, parametric features were normalised according to their majority distribution in order to share an equal range with non-parametric features:

$$Norm(v) = \begin{cases} 1, v > UB(v) \\ \dfrac{v - LB(v)}{UB(v) - LB(v)}, LB(v) < v < UB(v) \\ 0, v < LB(v) \end{cases} \quad (1)$$

where $LB(v)$ and $UB(v)$ are the lower and upper bound representing the 5^{th} and 95^{th} percentile within the target feature values. Next, we used an SVM-based selection scheme with the evaluation of area under the receiver operating characteristic curve (AUC) to optimise the feature set for the general case (**Fig 3**). In this scheme, we introduced the Fisher-Markov Selector [81] to calculate the importance of an individual feature. We assumed that the usage of better

BEGIN

Initialisation: Balanced dataset $S_0 = \{(L_1, v_1^0), .., (L_1, v_n^0), (L_2, v_{n+1}^0), \ldots, (L_2, v_{2n}^0)\}$, original feature set $F_m = (f_1, f_2, \ldots, f_m)$, machine learning classifier C, feature evaluation algorithm A, prediction evaluation criterion E, loop pointer $i = 2$.

(1) Evaluate the importance of individual feature $a = A(S_o)$.

(2) Create descending rank list based on the feature importance in a, $L = (f_1', f_2', \ldots, f_m')$.

(3) Use the most important feature to create feature set $F_1 = f_1'$.

(4) Update feature vector v_x^1, dataset S_1 and evaluate the prediction, $P_1 = C(S_1)$, $e_1 = E(P_1)$.

While $i \leq m$:

(5) Update feature set to include one well-performed feature based on L, $F_i = (f_1', \ldots, f_i')$;

(6) Update feature vector v_x^i, dataset S_i and evaluate the prediction, $P_i = C(S_i)$, $e_i = E(P_i)$;

(7) Calculate the improvement after including the new feature, $I_i = (e_i - e_{i-1})/e_{i-1}$;

(8) Update loop pointer $i = i + 1$.

End

Output: F_i achieving the best e_i and F_i achieving the best I_i.

END

Fig 3. The pseudo-code of the feature selection Scheme 1. We used the SVM model [51] as the base machine learning classifier and the Fisher-Markov Selector [81] to calculate the importance of an individual feature. AUC was chosen as the prime criterion to evaluate the prediction performance on datasets with multiple labels. Abbreviations: SVM, support vector machine; AUC, area under the receiver operating characteristic curve.

https://doi.org/10.1371/journal.pcbi.1009720.g003

performing features are less likely to negatively influence the complementarity of features in the set, which is crucial to training and modelling [40]. This feature selection scheme produced two outcomes: the optimum feature set and the lowest number of features.

The complementarity of different features implies information synergies, which can be measured by calculating the change of system entropy after the introduction of the additional features [40,82,83]. However, it is hard to discriminate if the combination of several random features can achieve better complementarity compared to using an equal number of well performing features. The selection strategy requires reconsideration if the impact of feature synergy has overwhelmed the usage of 'important' features on the prediction performance. Here, we use a second feature selection scheme which takes into account both feature importance and complementarity (**Fig 4**). As opposed to the first feature selection scheme (**Fig 3**), this scheme was processed by focusing on a set of features with good complementarity. It contained two main branches: the first expands the coverage of features by introducing well-performing features, while the second reduces the dimension of the feature sets by removing poorly performing features.

Performance evaluation

In order to assess the performance of different feature subsets, we adopted five-fold cross-validation on training datasets (dataset S1' and S2'), in which human proteins were randomly divided into five nearly equal parts and further generated five different testing (one portion) and training (the remaining four portions) sets. The overall quality of prediction models constructed from the feature subset was then evaluated based on the produced prediction scores via six criteria including sensitivity, specificity, accuracy, precision, Matthews Correlation Coefficient (MCC) [84] and AUC on the combination of five separate testing results. The evaluation of other independent testing datasets was also processed with the aforementioned six criteria, except in the case of the reference dataset S3, testing dataset S5 and S6, which only used sensitivity controlled by the threshold.

BEGIN

Initialisation: Feature sets with good complementarity $F_0' = (f_1, f_2, \ldots, f_s)$, the rest feature list $F_0 = (f_{s+1}, f_{s+2}, \ldots, f_m)$, balanced dataset $S_0 = \{(L_1, v_1^0), .., (L_1, v_n^0), (L_2, v_{n+1}^0), \ldots, (L_2, v_{2n}^0)\}$, machine learning classifier C, feature evaluation algorithm A, prediction evaluation criterion E, loop pointer $i = j = 1$.

(1) Evaluate the importance of individual feature $a = A(S_o)$.

(2) Create descending rank list for F_0 based on the feature importance in a, $L_1 = (f_1', f_2', \ldots, f_{m-s}')$.

(3) Create ascending rank list for F_0' based on the feature importance in a, $L_2 = (f_1'', f_2'', \ldots, f_s'')$.

While $i \leq m - s$:

(4) Update feature set to include one well-performed features based on L_1, $F_i = (f_1, \ldots, f_s, f_1', \ldots, f_i')$;

(5) Update feature vector v_x^i, dataset S_i and evaluate the prediction, $P_i = C(S_i)$, $e_i = E(P_i)$;

(6) Update loop pointer $i = i + 1$.

End

(7) Determine F_i achieving the best e_i.

While $j < s$:

(8) Update feature set to remove one badly-performed feature based on L_2, $F_j' = F_i - (f_1'', \ldots, f_j'')$;

(9) Update feature vector v_x^j, dataset S_j and evaluate the prediction, $P_j = C(S_j)$, $e_j = E(P_j)$;

(10) Update loop pointer $j = j + 1$.

End

Output: F_j' achieving the best e_j.

END

Fig 4. The pseudo-code of the feature selection Scheme 2. We used the SVM model [51] as the base machine learning classifier and the Fisher-Markov Selector [81] to calculate the importance of an individual feature. AUC was chosen as the prime criterion to evaluate the prediction performance on datasets with multiple labels. Abbreviations: SVM, support vector machine; AUC, area under the receiver operating characteristic curve.

https://doi.org/10.1371/journal.pcbi.1009720.g004

Results

HIV-1-host interaction pairs

Compared with available benchmark datasets [22–24,27,29], our main dataset S1 includes more HIV-1-host PPI data than previous studies (**Fig 5A**). The majority of HIV-1-host molecular interactions are associated with *env*-mediated membrane fusion [1–3] and *tat*-mediated transcellular transport [4,5,85]. In our main dataset S1, there are 996 (~35%) VIPs with interactions with multiple HIV-1 proteins. Some VIPs such as nuclear factor kappa B subunit 1 (NFKB1), interferon gamma (IFNG) and interferon beta 1 (IFNB1) are reported to interact with products produced by almost all HIV-1 genes [86–92]. **Fig 5B** illustrates the preference of co-occurring HIV-1-host PPIs interfering or being induced by the same VIP. It reveals a picture of host targets shared among HIV-1 gene products of *tat*, *env*, *gag*, *nef*, *gag-pol* and *vpr*, and the interaction preference underlying HIV-1 invasion, replication and assembly [1,7,9]. Despite the rank being ordered by the number of HIV-1-host interacting pairs (**Fig 5A**), HIV-1 *tat*-interacting proteins were marginally more frequently connected to *vpr* than *gag-pol*: 0.16 versus 0.14 per VIP, respectively. Interestingly, *tat* was less involved in the interactions with HIV-1 *gag*-, *nef*- and *gag-pol*-interacting proteins than expected. HIV-1 *env*-interacting proteins showed a preference to interact with *nef*, which is also involved in the early stage of the HIV-1 infection [7].

Statistical results indicate an overlap of human host proteins targeted by *env*, *tat* and *nef*. Forward VIPs targeted by HIV-1 *gag-pol*, *vif* and *rev* were less likely to interact with other

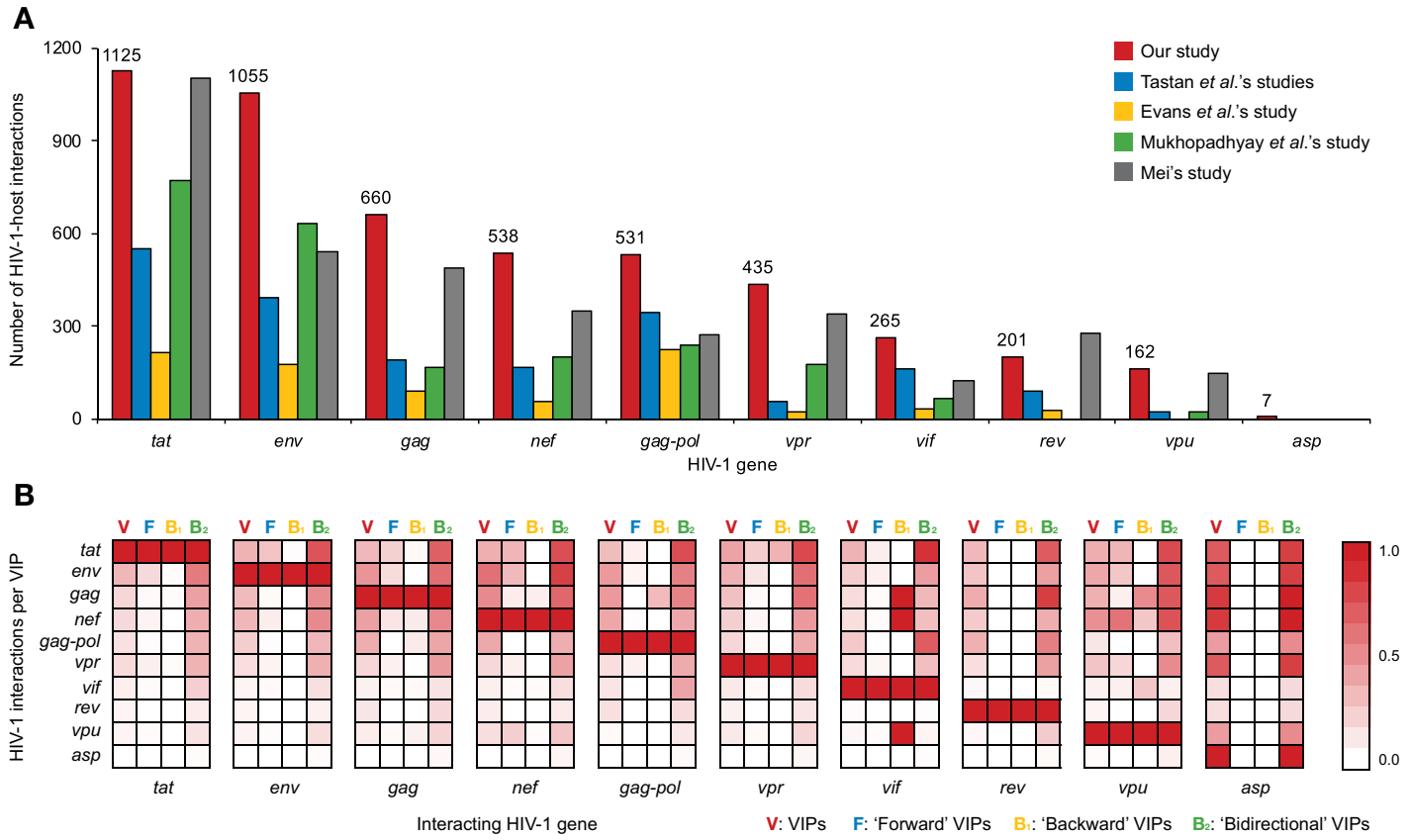

Fig 5. Comparison of HIV-1-host interaction datasets. (A) Comparison of datasets used in previous studies [22–24,27,29]. (B) Illustration of the preference of co-occurring HIV-1-host interactions for the VIPs (**V**), forward VIPs (**F**), backward VIPs ($\mathbf{B_1}$) and bidirectional VIPs ($\mathbf{B_2}$). Abbreviations: HIV-1, human immunodeficiency virus type 1; VIP, HIV-1 interacting human protein.

https://doi.org/10.1371/journal.pcbi.1009720.g005

HIV-1 proteins. An estimated 61% of forward VIPs targeted by *vpu* were also influenced by *nef*. After checking data with more detailed directionality information, we found that the forward or backward VIPs were generally associated with fewer interactions, while the bidirectional VIPs tended to be associated with higher numbers of interactions (**Figs 5B and S1**). Compared with other forward VIPs, *nef*- or *vpu*-interacting forward VIPs tended to be targeted by more HIV-1 proteins (Mann–Whitney U test: P = 3.0E-35). Meanwhile, *vpu*-interacting backward VIPs were more frequently targeted by multiple HIV-1 proteins than other backward VIPs (P = 4.2E-05). Collectively, it is common to observe human host proteins interacting with multiple HIV-1 proteins [44].

Feature analyses of the compiled data

In this study, we obtained 2881 nonredundant VIPs from 16215 HIV-1-host PPI records and 7261 high-quality non-VIPs from the human proteome. 1530 (~53%) of the VIPs (datasets S2 and S3) showed clear directionality: forward, backward, or bidirectional (Fig 2). In total, 671 features were collected from the genetic sequence, proteomic sequence, annotation and interaction profile data. To investigate the predictive signals in our compiled data, we analysed the feature representation in different human proteins.

Our characterisation using evolutionary-related features revealed a consistent pattern linked to HIV-1 infection (S1 Appendix). Higher numbers of protein-coding transcripts,

duplication rates and evolutionary conservation correlate with the HIV-1-host PPIs. A significantly biased distribution of one-transcript or one-protein-coding-transcript human genes in the VIP versus non-VIP classes provided a strong signal of inhibition associated with the HIV-1 infection (**S1 Appendix**, Pearson's Chi-squared test: M_1 = 12.4%, M_2 = 29.3%, P = 2.8E-71). These results suggest that high conservation has a tendency to be associated with pro-viral interactions, consistent with the documented ancient nature of virus-host interactions [93].

Analyses of the nucleotide and protein sequences were conducted using 354 features. The results indicated that VIPs and non-VIPs had some significant differences in their sequence patterns, e.g., in nucleobase composition linked to putative SLiMs (**S2 Appendix**). For instance, enrichment of adenine, depletion of cytosine and differential codon usage preferences of VIP genes influenced the distribution of amino acids in the protein sequence [94], which also contributed to the signal distinguishing the VIPs from non-VIPs. 85 SLiMs were detected to be more enriched in VIP sequences than in non-VIP sequences (Pearson's Chi-squared test: $P<0.05$). Co-occurrence of these putative SLiMs showed a cumulative effect resulting in better classifying of VIPs versus non-VIPs. Pro-viral and pro-host signatures of the VIPs were also reflected in sequence patterns and by intrinsic disorder status in the protein sequence. These results demonstrate the differential representation of sequence-based features provides a useful signal for prediction purposes.

Distinct from the aforementioned evolutionary-related and sequence-based features, annotation-based features are more straightforward with direct functional relevance [95]. Analyses of GO profiles revealed that the VIPs were more involved in cellular process (GO:0009987), binding (GO:0005488) and had a stronger association with organelles in the host cell (GO:0043226) than non-VIPs (Pearson's Chi-squared test: P = 9.9E-118, 1.0E-84 and 2.0E-70, respectively) (**S3 Appendix**). Within the group of VIPs, the bidirectional VIPs were highlighted for their prevalent response to stimulus (GO:0050896) and frequent involvement in biological regulation (GO:0065007) (**S3 Data**). Analyses of tissue tropisms indicated that the VIPs were more likely to be found in heart- or hematopoietic system-related tissue (**S3 Appendix**). Compared with the forward VIPs, the backward VIPs were less involved in the hematopoietic system but were more expressed in brain-related tissues, such as the brain stem and cerebral lobe (**S3 Data**). Cells originating from stem cells and differentiating in lymphoid tissues were favoured by backward VIPs and the relationship between backward VIPs and CD8+-presenting cells was evident, showing a clear relationship between HIV-1 infection and the host antiviral immune responses [96,97]. In view of the nature of the HIV-1-host molecular interactions, these function-related features are anticipated to perform well in the prediction tasks. However, they may not represent an optimum property in machine learning tasks due to gaps in annotations.

Performance of different feature sets in the training stage

Models for predicting the VIPs. In this study, we encoded 537 features for the prediction of the VIPs. According to the data source from which they were extracted, we divided these features into four categories: genome-based, proteome-based, annotation-based and interaction profile-based features. We first tested the performance of features in different categories on the balanced training datasets and found that annotation-based features performed the best, achieving the highest AUC value at 0.8090 on dataset S1' (**Table 2**). On the same dataset, the combination of interaction profile-based features produced some good predictions even if only 10 features were included. However, the performance of proteome-based features was poor on dataset S1'. By combining all of the encoded 537 features, we found the classifier could produce a better performance (AUC = 0.8324) than using features by individual

Table 2. The performance of different feature sets on the training datasets over five-cross validations.

Dataset[a]	Algorithm	Features	Features number	Threshold[e]	Sensitivity	Specificity	Accuracy	MCC	AUC
S1'	SVM	Genetic sequences	107	0.51	0.613	0.700	0.656	0.314	0.7118
	SVM	Proteomic sequences	128	0.51	0.595	0.649	0.622	0.244	0.6641
	SVM	Annotations	292	0.57	0.663	0.806	0.735	0.475	0.8090
	SVM	Interaction profiles	10	0.52	0.611	0.777	0.694	0.394	0.7487
	SVM	Combination	537	0.56	0.690	0.817	0.754	0.512	0.8324
	KNN[b]	Combination	537	0.35~0.39	0.766	0.633	0.699	0.402	0.7772
	DT[c]	Partial	278	N/A	0.633	0.642	0.637	0.275	N/A
	RF[d]	Random	Random	0.44~0.52	0.733±0.035	0.752±0.030	0.742±0.004	0.486±0.009	0.8157±0.0031
	SVM	Top-ranked 33	33	0.54	0.645	0.718	0.681	0.363	0.7468
	SVM	Top-ranked 193	193	0.48	0.748	0.751	0.750	0.499	0.8261
	KNN[b]	Optimum	441	0.43~0.48	0.689	0.720	0.705	0.410	0.7734
	SVM	Optimum	441	0.52	0.727	0.787	0.757	0.514	0.8344
S2'	SVM	Genetic sequences	107	N/A[f]	N/A[f]	N/A[f]	N/A[f]	N/A[f]	N/A[f]
	SVM	Proteomic sequences	164	0.40	0.860	0.633	0.747	0.507	0.8023
	SVM	Annotations	292	0.46	0.767	0.520	0.643	0.296	0.6786
	SVM	Interaction profiles	21	0.51	0.740	0.633	0.687	0.375	0.7108
	SVM	Combination	584	0.46	0.807	0.553	0.680	0.372	0.7383
	KNN[b]	Combination	584	0.50~0.54	0.400	0.833	0.617	0.259	0.6501
	DT[c]	Partial	66	N/A	0.673	0.660	0.667	0.333	N/A
	RF[d]	Random	Random	0.38~0.58	0.706±0.134	0.710±0.167	0.708±0.030	0.432±0.045	0.7609±0.0270
	KNN[b]	Optimum	129	0.27~0.36	0.487	0.873	0.680	0.390	0.7509
	SVM	Optimum	129	0.44	0.853	0.680	0.767	0.542	0.8260

[a]Dataset S1' and S2' were balanced training datasets constructed via an undersampling strategy [70] from dataset S1 and S2, respectively (Table 1). Compositions of these two datasets are provided in S2 Data.

[b]k-value here was determined as the square root of the size of the training samples in the five-fold cross validation

[c]the DT algorithm selected 278 and 66 features from the original feature sets for the two modelling tasks

[d]the RF algorithm used 50 randomly grown trees and the modelling and validation procedures were repeated 10 times

[e]threshold was set by maximizing the value of MCC

[f]'N/A' was denoted if the prediction quality of the generated classifier was worse than a random guess.

Abbreviations: SVM, support vector machine; KNN, k-nearest neighbors; DT, decision tree; RF, random forest; MCC, Matthews Correlation Coefficient; AUC, area under the receiver operating characteristic curve.

https://doi.org/10.1371/journal.pcbi.1009720.t002

categories (AUC = 0.7118, 0.6641, 0.8090, 0.7487, respectively) (Table 2). We compared the SVM with another three machine learning models: KNN, DT and RF [33]. We used the square root of the size of the training samples as the k-value for the KNN algorithm [98]. We found this algorithm was biased to the positive class and did not achieve a better prediction performance than the SVM model. The DT algorithm was designed with a feature selection scheme, which helped it to better split the dataset for lower system entropy [82]. It used 278 out of 537 features and had the worst performance among the different machine learning algorithms compared. We initialised the RF algorithm with 50 trees and repeated the modelling process ten times to balance bootstrapping of the dataset and selection of features [99]. The prediction performance of the RF algorithm on S1' was promising but did not surpass that of the SVM model. These results suggest that the majority of features encoded for the prediction of the VIPs are contributing to the signal, but the complete feature set is not optimal for reliable prediction since it includes some poorly performing features.

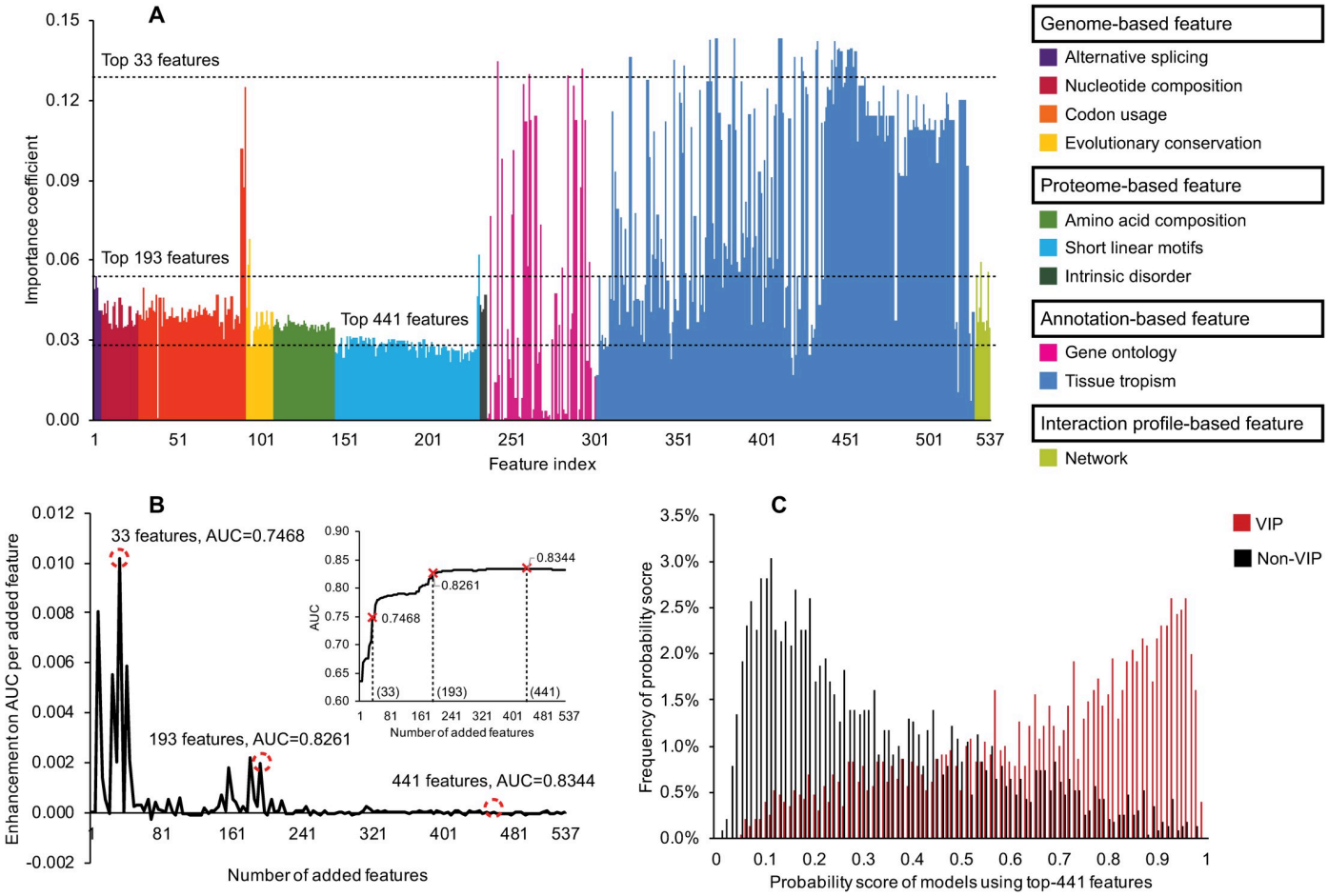

Fig 6. The performance of different features for the prediction of VIPs. (A) The importance of different features. (B) Enhancement the prediction performance by adding more features. (C) The distribution of prediction scores (for VIPs and non-VIPs) generated by model using the top 441 features. In (A) the importance of an individual feature was recorded by averaging the results on the balanced training datasets generated by ten-round undersampling procedures [70] on dataset S1. The ranked list of the encoded 537 features is provided in **S4 Data**. Abbreviations: HIV-1, human immunodeficiency virus type 1; VIPs, HIV-1 interacting human proteins; non-VIPs, non-HIV-1 interacting human proteins; AUC, area under the receiver operating characteristic curve.

https://doi.org/10.1371/journal.pcbi.1009720.g006

In order to find a better feature subset for the prediction of the VIPs, we first used the Fisher-Markov Selector [81] to calculate the importance of individual features (**Fig 6A**). The results demonstrated the importance of gene ontology (e.g., involvement in metabolic process, ranked 24th) and tissue tropism features for prediction (e.g., expression in monocyte, ranked 1st), even if they were used individually. Based on this ranking of individual features, we then used our first feature selection strategy (**Fig 3**) and five-cross validations to optimise the prediction model. **Fig 6** shows that the classifier obtained reasonable prediction with fairly low numbers of features, e.g., the top 33 features (**S2A Fig**) are near the first inflection point (**Fig 6B**), while the subset of 193 features were considered optimal as they achieved decent prediction performance (AUC = 0.8261, **S2B Fig**) and there will be less issues with missing data or errors in annotations [100]. The subset of 441 features maximised the prediction performance and are comprised of 105 genome-based features, 84 proteome-based features, 243 annotation-based features and nine interaction profile-based features. The distribution of prediction scores for VIPs and non-VIPs was negatively and positively skewed, with most values clustered around the right and left tails of the distribution, respectively (**Fig 6C**).

Models for predicting the direction of the HIV-1-host PPIs. We encoded 584 features for the VIPs to predict pro-host versus pro-virus directionality in the HIV-1-host molecular interactions. Predictions on the training dataset S2' were different from those on dataset S1'. The performance of genome-based features was even worse than a random prediction (Table 2). The combination of proteomic features produced a highly predictive model. The performance of annotation features was not as good as anticipated and the combination of all features made the classifier worse than only using proteomic features. This indicates a big difference between the two prediction tasks highlighted in this study. After checking the importance of features with the Fisher-Markov Selector [81] we found the difference between the generated importance scores was not obvious (**S3 Fig and S4 Data**), which meant the contribution of individual features to the prediction model had not changed appreciably. The comparison of results from the different machine learning algorithms demonstrated that the SVM classifier still worked on dataset S2' (Table 2). These results suggest that the overall complementarity of these 584 features is not as good as those used for predicting the VIPs. There may be a large number of noisy features involved in the complete set, which suppresses the performance of some feature combinations when using our first selection strategy (Fig 7A).

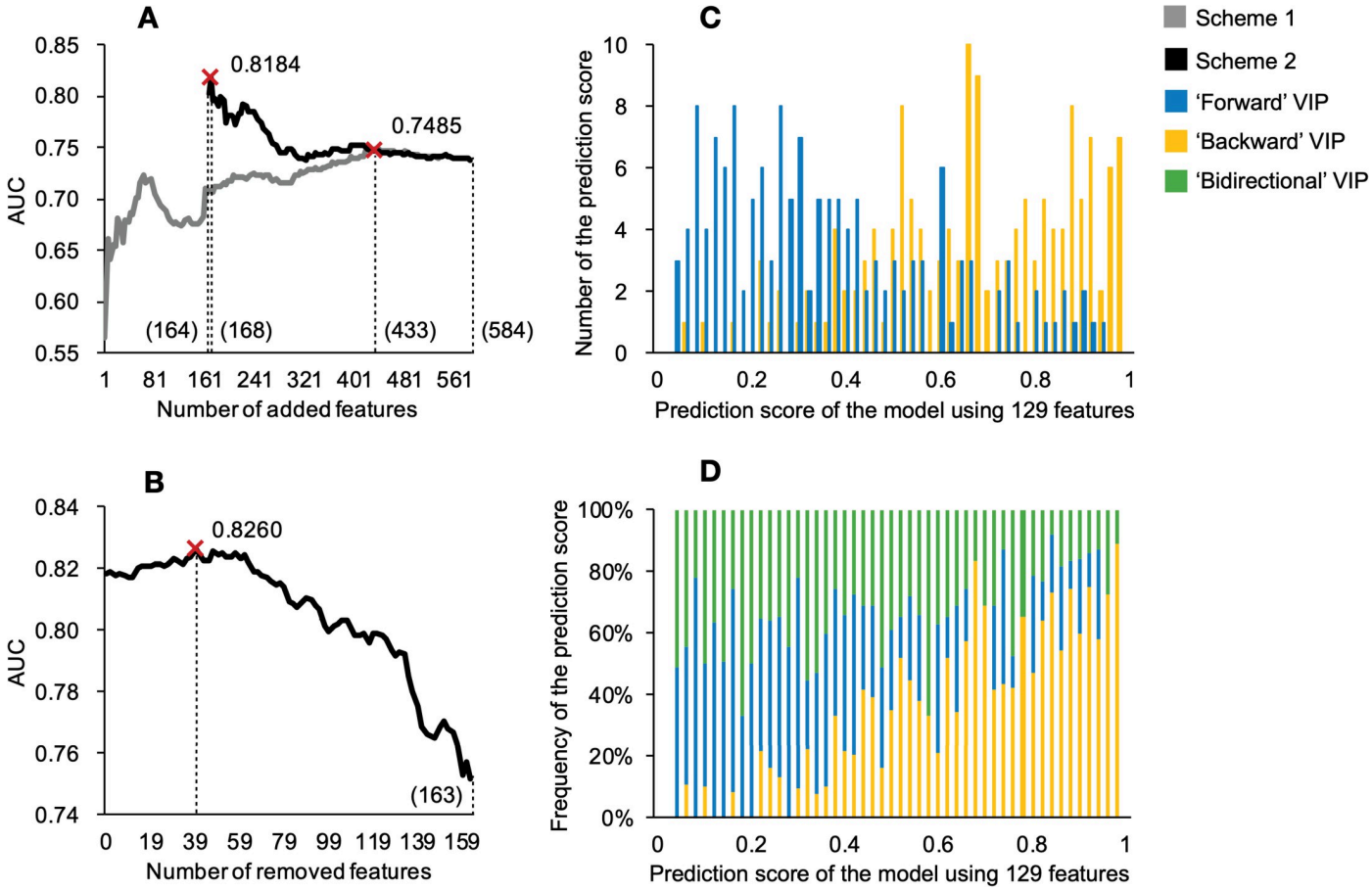

Fig 7. The performance of different features for predicting the backward and forward VIPs. (A) AUC values for increasing numbers of features. (B) AUC values for decreasing numbers of proteome-based features. (C) The counts of prediction scores (for pro-viral/forward VIPs and pro-host/backward VIPs) generated by model using 129 optimum features. (D) The percentage of forward, backward and bidirectional VIPs within different regions of prediction scores (scale = 0.02). Abbreviations: HIV-1, human immunodeficiency virus type 1; VIP, HIV-1 interacting human protein; AUC, area under the receiver operating characteristic curve.

https://doi.org/10.1371/journal.pcbi.1009720.g007

Thus, we applied our second feature selection strategy to optimise the prediction model (Fig 4). We assumed that the proteomic features might be an ideal set with good complementarity for the initialization as they were better-performing than features in the other categories (Table 2). We found performance of the classifier was enhanced, however, it started to decrease after adding four non-proteomic sequence features. We, thus, removed the poorly-performing features in the proteomic sequence feature set. We identified an optimum model generated by 129 features (Fig 7B). In that feature subset, 36 amino acid composition, 85 SLiM, four intrinsic disorder, two gene ontology and two tissue tropism features were included (S4 Data). Compared with the model generated from the complete feature set, the model using 129 optimum features enhanced the performance by more than 10% from the perspective of the AUC (Table 2). Likewise, on the training dataset S2', the SVM was still superior to the KNN, DT and RF algorithms. Additionally, the model generated from all proteomic features was also recommended as it only required the information from the protein sequence to make reasonable predictions (S4 Fig).

Interestingly, testing on the reference dataset S3 suggests that the bidirectional VIPs are closer in properties to the forward VIPs than to the backward VIPs. The forward VIPs may be 'responding' to the HIV-1 infection and target HIV-1, making them 'bidirectional' [101]. The backward VIPs are less likely to be targeted by HIV-1 so their chances of becoming 'bidirectional' are relatively low. The recommended direction based on the prediction score generated by the model using 129 optimum features is listed in S1 Table. We could confidently label 60% of the generated VIPs based on the prediction scores as backward, forward, or bidirectional. For prediction scores located in specific ranges, our confidence on the direction of the HIV-1-host molecular interactions could reach as high as 89%.

Performance on the testing datasets

In this study, we produced three models with the top-33, top-193 and top-441 features on the whole training dataset S1' for the prediction of VIPs, namely PreVIP-33, PreVIP-193 and PreVIP-441, respectively. Independent testing datasets prepared to assess the generalization capability of these three models was derived from our main dataset S1 through an undersampling strategy [70]. They consist of a random set of 577 VIPs and 4957 non-VIPs. The imbalance ratio of positives (VIPs) to negatives (non-VIPs) in this testing dataset is close to 1:8. PreVIP-33 could successfully predict 40.9% of VIPs and 88.1% of non-VIPs under a threshold of 0.73. The corresponding AUC value of PreVIP-33 was 0.7323. Under the same threshold, the sensitivity and specificity of PreVIP-193 increased to 45.6% and 91.2%, respectively. The optimum threshold for PreVIP-193 was 0.82, under which 34.7% of the VIPs and more than 95% of the non-VIPs were correctly predicted. Among the generated three models, PreVIP-441 achieved the best performance with an AUC value of 0.8079, with the performance of PreVIP-193 close to this (AUC of 0.8034) (Table 3). In contrast, PreVIP-33 did not perform well on the testing dataset S1". When attempting to successfully predict more than half of the VIPs, the ratios of false positives produced by PreVIP-441, PreVIP-193 and PreVIP-33 were 10%, 11% and 19%, respectively.

As for predicting the direction of the HIV-1-host molecular interactions, we generated two models with the optimal 129 features and the overall 164 proteomic sequence features on the whole training dataset S2', namely PreDIR-129 and PreDIR-164. An independent testing dataset prepared to assess the generalization capability of these two models was derived from our main dataset S2 using an undersampling strategy [70]. It is comprised of a random 38 VIPs and 857 non-VIPs. The imbalance ratio of positives (backward VIPs) to negatives (forward VIPs) in this testing dataset is close to 1:22. Compared with PreDIR-129, PreDIR-164 was

Table 3. The performance of features with different categories on the testing datasets.

Dataset	Model	Feature source	Threshold[a]	Sensitivity	Specificity	Accuracy	Precision	MCC	AUC
S1"	PreVIP-33	Annotation	0.73	0.409	0.881	0.832	0.285	0.248	0.7323
	PreVIP-193	Multiple	0.82	0.347	0.959	0.895	0.495	0.359	0.8034
	PreVIP-441	Multiple	0.73	0.492	0.911	0.867	0.391	0.365	0.8079
S2"	PreDIR-164	Proteomic sequence	0.53	0.658	0.762	0.758	0.109	0.194	0.7110
	PreDIR-129	Multiple	0.70	0.474	0.873	0.856	0.142	0.200	0.7057
S5[b]	PreVIP-193	Multiple	0.82	Sensitivity = 0.577					
	PreVIP-193	Multiple	0.50	Sensitivity = 0.906					
	PreVIP-441	Multiple	0.73	Sensitivity = 0.701					
	PreVIP-441	Multiple	0.50	Sensitivity = 0.910					
S6[b]	PreVIP-193	Multiple	0.82	Sensitivity = 0.416					
	PreVIP-193	Multiple	0.50	Sensitivity = 0.806					
	PreVIP-441	Multiple	0.73	Sensitivity = 0.596					
	PreVIP-441	Multiple	0.50	Sensitivity = 0.817					

[a]thresholds on S1" and S2" were set by maximizing the value of MCC. On testing dataset S5 and S6, two thresholds, i.e., 0.82 and 0.73 were set according to the best performance of PreVIP-193 and PreVIP-441 on testing dataset S1". In addition, a neutral threshold (0.5) was added for crude assessments.

[b]prediction results on testing dataset S5 and S6 are provided in **S5 Data**.

Abbreviations: MCC, Matthews Correlation Coefficient; AUC, area under the receiver operating characteristic curve; HIV-1, human immunodeficiency virus type 1; VIPs, HIV-1 interacting human proteins; PreVIP-33, machine learning model generated from training dataset S1' with the top 33 features for the VIP prediction task; PreVIP-193, machine learning model generated from training dataset S1' with the top 193 features for the VIP prediction task; PreVIP-441, machine learning model generated from training dataset S1' with the optimum 441 features for the VIP prediction task; PreDIR-164, machine learning model generated from training dataset S2' with 164 proteome-based features for the directionality prediction task; PreDIR-129, machine learning model generated from training dataset S2' with the optimum 129 features for the directionality prediction task.

https://doi.org/10.1371/journal.pcbi.1009720.t003

generally a bit more powerful for achieving higher AUC, at 0.7110 (**Table 3**). The optimum threshold for PreDIR-129 was 0.70, under which 47.4% of the backward VIPs and 87.3% of the forward VIPs were successfully predicted. By contrast, to successfully filter the same number of negatives, PreDIR-164 produced three more false negatives, which showed its drawback in recognising forward VIPs.

To further assess the quality of PreVIP-193 and PreVIP-441, we introduced two testing datasets from the HIV-1 infection pathway in Reactome [60] (dataset S5) and an HIV-1 host-dependency epistasis map [61] (dataset S6). **Fig 8** shows that our prediction models performed well in recognising VIPs confirmed with experimental evidence. On testing dataset S5, PreVIP-441 could recognise 70.1% of VIPs under a threshold of 0.73, about 40% more than its expected performance (**Table 3**). It was also capable of successfully predicting more than 90% of VIPs when using a threshold of 0.5. Under the same threshold, PreVIP-193 achieved a similar performance as PreVIP-441. The improvement under the threshold of 0.82 reached as high as 66% when compared with its expected sensitivity (34.7%). Thus, these results demonstrate good generalization capabilities of our models on predicting VIPs involved in the host sub-systems hijacked during HIV-1 infection [60]. Their performance on the testing dataset S6 was also promising with an estimated 20% improvement. The prediction results of PreVIP-193 and PreVIP-441 on testing datasets S5 and S6 are shown in **S5 Data**.

On the blind testing dataset S4, we used PreDIR-129 to predict the direction tag for 1351 'Undefined' VIPs (**Fig 2**). According to known information about potential direction (**S1 Data**) and the recommendation stated in **S1 Table**, 511, 540 and 300 undefined VIPs were predicted as backward, forward and bidirectional, respectively (**S6 Data**). The prediction scores for the putative different VIPs showed significant differences in the majority of regions (**S5**

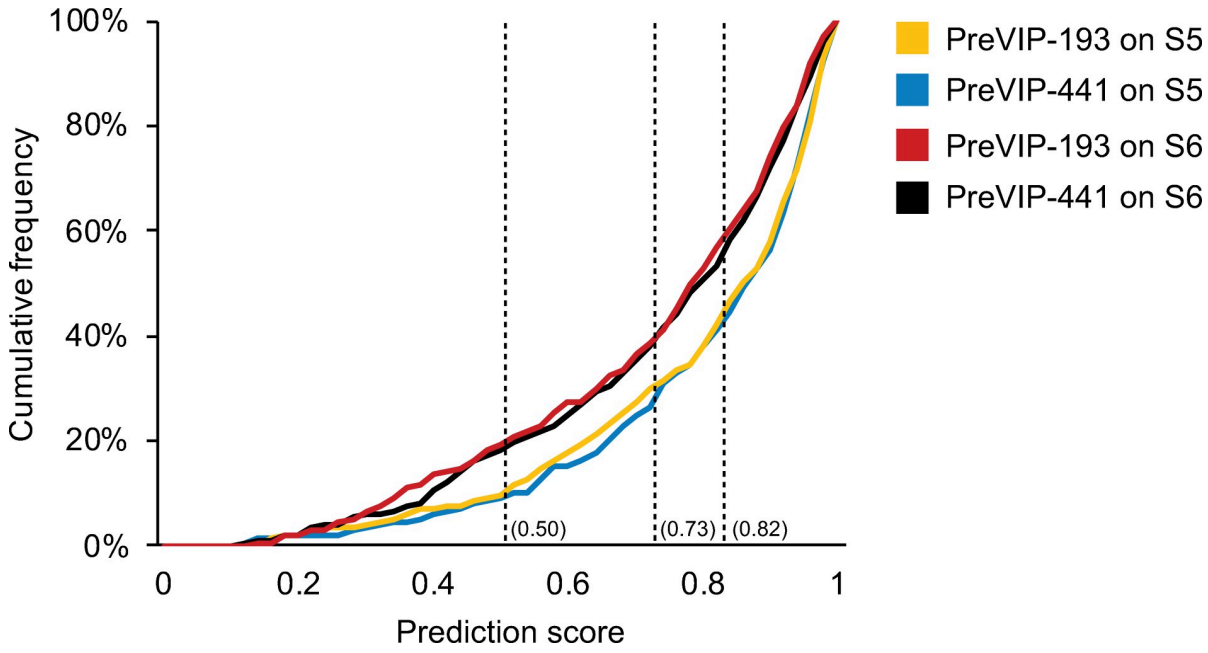

Fig 8. Cumulative distribution of prediction probabilities on the testing datasets S5 and S6. Dataset S5 and S6 were retrieved from Reactome [60] and Gordon *et al.*'s study [61] for the purpose of predicting VIPs. The composition of dataset S5 and S6 is provided in the **S2 Data** file. Prediction results on testing dataset S5 and S6 are provided in **S5 Data**. Abbreviations: VIPs, HIV-1 interacting human proteins; PreVIP-193, machine learning model generated from training dataset S1' with the top 193 features for the VIP prediction task; PreVIP-441, machine learning model generated from training dataset S1' with the optimum 441 features for the VIP prediction task.

https://doi.org/10.1371/journal.pcbi.1009720.g008

Fig) (Mann–Whitney U test: P = 2.5E-124, 2.0E-43, 4.9E-19, respectively). 66 of the undefined VIPs had a high probability of being 'backward' interactions and in terms of function the literature shows connections to brain-related diseases such as autosomal recessive neurodevelopmental disorder deficiency [102] and Huntington disease [103] (**S2 Table**). 63 of the undefined VIPs were very likely to be 'forward' interactions and were involved in some immune system pathways (**S2 Table**). 50 of the undefined VIPs showed strong signals of being 'bidirectional' interactions. Interestingly, they appear to be targets of other viruses like human papillomavirus [104] and hepatitis virus [105] (**S2 Table**).

Discussion

In this study, we propose an *in silico* approach to investigate HIV-1-host molecular interactions with a focus on prediction of the directionality of the virus-host interaction. We used the detailed curation of the biological nature of known HIV-1-host interactions in the HHID [44] to partition interactions as those required by the virus to manipulate the host molecular subsystems versus host responses to virus infection. Using this dataset, we design a predictive system in which human proteins can be quickly evaluated for their potential to target host (a host-dependency factor), be targeted (the antiviral response), or both (bidirectional interactions). A web server is available at http://hivpre.cvr.gla.ac.uk/. It supports six different identifiers for over 80000 human peptides and can carry out 1000 predictions in less than 15 seconds.

In previous studies [22–30], VIPs were usually labeled based on their interacting HIV-1 status only. According to the data we retrieved from the HHID [44], 1467 out of 3854 human proteins, including some key receptors (e.g., CD4 and CCR5), have interactions with protein products of different HIV-1 genes (**S1 Data**). Such multi-target issues can be accommodated

by integrating information on the molecular interactions between human proteins and the HIV-1 interaction type. Previously published prediction-based papers [22–24,26–30] have not accounted for the direction of the HIV-1-host molecular interaction. By contrast, our consideration of the interaction direction contributes to a better understanding of the HIV-1-host interactions and the discovery of potential drug targets [106].

Analysing evolutionary-related information in the transcriptomic and genomics data (**S1 Appendix**), we found HIV-1 was more likely to interact with human proteins encoded from genes with higher numbers of protein-coding transcripts, higher duplication rates and more evolutionary conserved. Presumably this is at least partly because the evolutionary rates for duplicate genes have a tendency to be negatively correlated with the number of paralogues [107], and virus-interacting molecules are often relatively evolutionary ancient [93]. We discovered 85 VIP-enriched putative SLiMs and 121 backward VIP-enriched SliMs from the proteomic sequence data (**S2 Appendix**). We hypothesise that there are some motifs in the sequence of VIPs mediating molecular interactions, making them more likely to target or be the target of HIV-1. Human proteins with longer sequences have a higher probability of including some predictive sequence patterns than those with short sequences. For example, there are over 14000 residues in the sequence of a non-VIP, namely mucin 16 (MUC16), but only 17 VIP-enriched SliMs were observed. However, this signal needs to be treated with caution especially when large numbers of VIP-enriched and backward VIP-enriched SliMs are both detected in the same non-VIP sequence, e.g., midasin (MDN1) (n = 41 and 57). Such 'non-VIPs' may potentially be false negatives if some of their SLiM-enriched regions could interact with HIV-1 [108]. We obtained 225 experimentally verified tissue entries from the TISSUES database [48] but found some non-independence of features due to the hierarchical nature of this type of data(**S3 Appendix**). Nonetheless, the annotation data of tissue tropisms was sufficient for distinguishing VIPs from non-VIPs (**S6 Fig**). The later analysis also demonstrated the practical effectiveness of considering these features individually or in combination (**Fig 6A** and **Table 2**).

After finishing all prediction tasks, we assumed that some false negatives were still included in our dataset since we found some of the testing non-VIPs obtained very high prediction scores (**S3 Table**). Based on the testing result given by PreVIP-441 and PreVIP-193, 16 labelled non-VIPs might actually interact with HIV-1 proteins. For example, we found that adapter molecule crk (CRK), TGF-beta-activated kinase 1 and MAP3K7-binding protein 1 (TAB1) and interleukin-1 receptor-associated kinase 4 (IRAK4) are involved in the HIV-1 infection in the Kyoto Encyclopedia of Genes and Genomes (KEGG) database [109] but were not included in the HHID [44]. This provided further support for the predictive value of our machine learning approach. Some features of these human proteins also hinted at their possible roles as VIPs. For instance, alpha-synuclein (SNCA) had a high number of polymorphisms, contained 15 VIP-enriched SLiMs within its 140-length proteomic sequence, expressed in many VIP-preferred tissues and was highly connected with a degree of 168 in our constructed network [50]. As for the prediction of the interaction directionality, some results in **S1 Table** might be ambiguous when being used individually but higher confidence could be obtained when combining the information on known interaction directionality in **S1 Data**. For instance, elongin-B (ELOB) had a prediction score of 0.14 from PreDIR-129 so was initially predicted to be a forward VIP (**S1 Table**). However, since we found 18 records on the molecular interactions between ELOB and HIV-1 proteins and some outcomes of the interactions showed the clear direction of 'backward', ELOB is probably 'bidirectional' rather than only 'forward' acting.

In conclusion, reliably predicting HIV-1-host molecular interactions is a difficult task and to improve requires a better framework for understanding the nuances of the virus-host

relationship. Here we have introduced the directionality of the interaction to this task and demonstrated that there is a predictive signal embedded in the different types of molecules. Many of the features used, however, are only superficially capturing the information embedded in the molecules involved. We are confident that better training datasets and continued development of feature representation of molecules, for example, integrating protein structure and molecular interaction data, will lead to improved predictions in the near future.

Supporting information

S1 Appendix. Characterisation of features linked to alternative splicing and evolution.
(PDF)

S2 Appendix. Characterisation of features in different sequence patterns.
(PDF)

S3 Appendix. Characterisation of features in annotation and network profiles.
(PDF)

S1 Fig. The preference of co-occurring HIV-1-host interactions for different VIPs. Boxes in the plot represent the major distribution of values (from the first to the third quartile); outliers were added for values higher than two-fold of the third quartile; the cross symbol marks the position of the average value including the outliers; upper and lower whiskers showed the maximum and minimum values excluding the outliers. Abbreviations: HIV-1, human immunodeficiency virus type 1; VIP, HIV-1 interacting human protein.
(PDF)

S2 Fig. Prediction score generated by models using (A) the top-33 and (B) top-193 features on dataset S1' over five-cross validation. Abbreviations: HIV-1, human immunodeficiency virus type 1; VIPs, HIV-1 interacting human proteins; non-VIPs, non-HIV-1 interacting human proteins.
(PDF)

S3 Fig. Importance of individual features for predicting the backward and forward VIPs. Abbreviations: The importance score of individual features is recorded by averaging the results on the balanced training datasets generated by ten-round undersampling procedures on dataset S2. HIV-1, human immunodeficiency virus type 1; VIPs, HIV-1 interacting human proteins.
(PDF)

S4 Fig. Prediction score generated by models using proteomic features on dataset S2' over five-cross validation. Abbreviations: HIV-1, human immunodeficiency virus type 1; VIPs, HIV-1 interacting human proteins.
(PDF)

S5 Fig. Prediction score generated by PreDIR-129 on the blind testing dataset S4. Abbreviations: HIV-1, human immunodeficiency virus type 1; VIPs, HIV-1 interacting human proteins.
(PDF)

S6 Fig. Tissue tropisms for different human proteins. Abbreviations: HIV-1, human immunodeficiency virus type 1; VIPs, HIV-1 interacting human proteins; non-VIPs, non-HIV-1 interacting human proteins.
(PDF)

S1 Table. The recommended interaction direction for models using the 129 optimum feature set.
(PDF)

S2 Table. Undefined VIPs with very high probabilities indicating their directionality in the HIV-1-host molecular interactions.
(XLSX)

S3 Table. Non-VIPs with high prediction scores (>0.95) generated by PreVIP-441 and PreVIP-193.
(XLSX)

S1 Data. The list of HIV-1 interacting proteins.
(TXT)

S2 Data. The list of proteins sampled for training and testing.
(TXT)

S3 Data. Statistical data supporting results in the manuscript and appendix.
(XLSX)

S4 Data. The importance and usage of features in different machine learning models.
(TXT)

S5 Data. Prediction results on dataset S5 and S6.
(TXT)

S6 Data. Prediction results for VIPs with the undefined directions.
(TXT)

Acknowledgments

The authors wish to thank Prof Andrew Davison, Drs Suzannah Rihn, Ke Yuan, Vandana Ravindran and Sam Wilson for helpful discussions and recommendations, and Scott Arkison for help setting up the website.

Author Contributions

Conceptualization: Haiting Chai, Quan Gu, Joseph Hughes, David L. Robertson.

Data curation: Haiting Chai.

Formal analysis: Haiting Chai.

Funding acquisition: David L. Robertson.

Investigation: Haiting Chai.

Methodology: Haiting Chai.

Project administration: David L. Robertson.

Software: Haiting Chai.

Supervision: Quan Gu, Joseph Hughes, David L. Robertson.

Visualization: Haiting Chai.

Writing – original draft: Haiting Chai.

Writing – review & editing: Haiting Chai, Quan Gu, Joseph Hughes, David L. Robertson.

References

1. Brandenberg OF, Magnus C, Regoes RR, Trkola A. The HIV-1 entry process: a stoichiometric view. Trends Microbiol. 2015; 23(12): 763–774. https://doi.org/10.1016/j.tim.2015.09.003 PMID: 26541228
2. Lusic M, Siliciano RF. Nuclear landscape of HIV-1 infection and integration. Nat Rev Microbiol. 2017; 15(2): 69–82. https://doi.org/10.1038/nrmicro.2016.162 PMID: 27941817
3. Deeks SG, Overbaugh J, Phillips A, Buchbinder S. HIV infection. Nature reviews Disease primers. 2015; 1(1): 1–22. https://doi.org/10.1038/nrdp.2015.35 PMID: 27188527
4. Molle D, Maiuri P, Boireau S, Bertrand E, Knezevich A, Marcello A, et al. A real-time view of the TAR: Tat: P-TEFb complex at HIV-1 transcription sites. Retrovirology. 2007; 4(1): 1–5. https://doi.org/10.1186/1742-4690-4-36 PMID: 17537237
5. Debaisieux S, Rayne F, Yezid H, Beaumelle B. The ins and outs of HIV-1 Tat. Traffic. 2012; 13(3): 355–363. https://doi.org/10.1111/j.1600-0854.2011.01286.x PMID: 21951552
6. Malim MH, Emerman M. HIV-1 accessory proteins—ensuring viral survival in a hostile environment. Cell Host Microbe. 2008; 3(6): 388–398. https://doi.org/10.1016/j.chom.2008.04.008 PMID: 18541215
7. Seelamgari A, Maddukuri A, Berro R, de la Fuente C, Kehn K, Deng L, et al. Role of viral regulatory and accessory proteins in HIV-1 replication. Front Biosci. 2004; 9(9): 2388–2413. https://doi.org/10.2741/1403 PMID: 15353294
8. Balachandran A, Wong R, Stoilov P, Pan S, Blencowe B, Cheung P, et al. Identification of small molecule modulators of HIV-1 Tat and Rev protein accumulation. Retrovirology. 2017; 14(1): 1–21. https://doi.org/10.1186/s12977-016-0324-3 PMID: 28086923
9. Freed EO. HIV-1 assembly, release and maturation. Nature Reviews Microbiology. 2015; 13(8): 484–496. https://doi.org/10.1038/nrmicro3490 PMID: 26119571
10. Meyerson NR, Rowley PA, Swan CH, Le DT, Wilkerson GK, Sawyer SL. Positive selection of primate genes that promote HIV-1 replication. Virology. 2014; 454: 291–298. https://doi.org/10.1016/j.virol.2014.02.029 PMID: 24725956
11. Towers GJ, Noursadeghi M. Interactions between HIV-1 and the cell-autonomous innate immune system. Cell Host Microbe. 2014; 16(1): 10–18. https://doi.org/10.1016/j.chom.2014.06.009 PMID: 25011104
12. Valera M-S, de Armas-Rillo L, Barroso-González J, Ziglio S, Batisse J, Dubois N, et al. The HDAC6/APOBEC3G complex regulates HIV-1 infectiveness by inducing Vif autophagic degradation. Retrovirology. 2015; 12(1): 1–26. https://doi.org/10.1186/s12977-015-0181-5 PMID: 26105074
13. Shoji-Kawata S, Zhong Q, Kameoka M, Iwabu Y, Sapsutthipas S, Luftig RB, et al. The RING finger ubiquitin ligase RNF125/TRAC-1 down-modulates HIV-1 replication in primary human peripheral blood mononuclear cells. Virology. 2007; 368(1): 191–204. https://doi.org/10.1016/j.virol.2007.06.028 PMID: 17643463
14. Doyle T, Goujon C, Malim MH. HIV-1 and interferons: who's interfering with whom? Nat Rev Microbiol. 2015; 13(7): 403–413. https://doi.org/10.1038/nrmicro3449 PMID: 25915633
15. MacPherson JI, Dickerson JE, Pinney JW, Robertson DL. Patterns of HIV-1 protein interaction identify perturbed host-cellular subsystems. PLoS Comput Biol. 2010; 6(7): e1000863. https://doi.org/10.1371/journal.pcbi.1000863 PMID: 20686668
16. Engelman A, Cherepanov P. The structural biology of HIV-1: mechanistic and therapeutic insights. Nat Rev Microbiol. 2012; 10(4): 279–290. https://doi.org/10.1038/nrmicro2747 PMID: 22421880
17. Dorr P, Westby M, Dobbs S, Griffin P, Irvine B, Macartney M, et al. Maraviroc (UK-427,857), a potent, orally bioavailable, and selective small-molecule inhibitor of chemokine receptor CCR5 with broad-spectrum anti-human immunodeficiency virus type 1 activity. Antimicrobial Agents and Chemotherapy. 2005; 49(11): 4721–4732. https://doi.org/10.1128/AAC.49.11.4721-4732.2005 PMID: 16251317
18. Pinney JW, Dickerson JE, Fu W, Sanders-Beer BE, Ptak RG, Robertson DL. HIV–host interactions: a map of viral perturbation of the host system. AIDS. 2009; 23(5): 549–554. https://doi.org/10.1097/QAD.0b013e328325a495 PMID: 19262354
19. Dickerson JE, Pinney JW, Robertson DL. The biological context of HIV-1 host interactions reveals subtle insights into a system hijack. BMC Syst Biol. 2010; 4(1): 1–13. https://doi.org/10.1186/1752-0509-4-80 PMID: 20529270
20. Chen K-C, Wang T-Y, Chan C-h. Associations between HIV and human pathways revealed by protein-protein interactions and correlated gene expression profiles. PLoS One. 2012; 7(3): e34240. https://doi.org/10.1371/journal.pone.0034240 PMID: 22479575
21. Chen L, Keppler OT, Schölz C. Post-translational modification-based regulation of HIV replication. Front Microbiol. 2018; 9: 2131. https://doi.org/10.3389/fmicb.2018.02131 PMID: 30254620

22. Tastan O, Qi Y, Carbonell JG, Klein-Seetharaman J. Prediction of interactions between HIV-1 and human proteins by information integration. Biocomputing 2009: World Scientific; 2009. p. 516–527. PMID: 19209727

23. Qi Y, Tastan O, Carbonell JG, Klein-Seetharaman J, Weston J. Semi-supervised multi-task learning for predicting interactions between HIV-1 and human proteins. Bioinformatics. 2010; 26(18): i645–i652. https://doi.org/10.1093/bioinformatics/btq394 PMID: 20823334

24. Mukhopadhyay A, Maulik U, Bandyopadhyay S. A novel biclustering approach to association rule mining for predicting HIV-1–human protein interactions. PLoS One. 2012; 7(4): e32289. https://doi.org/10.1371/journal.pone.0032289 PMID: 22539940

25. Mukhopadhyay A, Ray S, Maulik U. Incorporating the type and direction information in predicting novel regulatory interactions between HIV-1 and human proteins using a biclustering approach. BMC Bioinformatics. 2014; 15(1): 1–22. https://doi.org/10.1186/1471-2105-15-26 PMID: 24460683

26. Dyer MD, Murali T, Sobral BW. Supervised learning and prediction of physical interactions between human and HIV proteins. Infect Genet Evol. 2011; 11(5): 917–923. https://doi.org/10.1016/j.meegid.2011.02.022 PMID: 21382517

27. Mei S. Probability weighted ensemble transfer learning for predicting interactions between HIV-1 and human proteins. PLoS One. 2013; 8(11): e79606. https://doi.org/10.1371/journal.pone.0079606 PMID: 24260261

28. Doolittle JM, Gomez SM. Structural similarity-based predictions of protein interactions between HIV-1 and Homo sapiens. Virol J. 2010; 7(1): 1–15. https://doi.org/10.1186/1743-422X-7-82 PMID: 20426868

29. Evans P, Dampier W, Ungar L, Tozeren A. Prediction of HIV-1 virus-host protein interactions using virus and host sequence motifs. BMC Med Genomics. 2009; 2(1): 1–13. https://doi.org/10.1186/1755-8794-2-27 PMID: 19450270

30. Becerra A, Bucheli VA, Moreno PA. Prediction of virus-host protein-protein interactions mediated by short linear motifs. BMC Bioinformatics. 2017; 18(1): 1–11. https://doi.org/10.1186/s12859-016-1414-x PMID: 28049414

31. Nourani E, Khunjush F, Durmuş S. Computational approaches for prediction of pathogen-host protein-protein interactions. Front Microbiol. 2015; 6: 94. https://doi.org/10.3389/fmicb.2015.00094 PMID: 25759684

32. Durmuş S, Çakır T, Özgür A, Guthke R. A review on computational systems biology of pathogen–host interactions. Front Microbiol. 2015; 6: 235. https://doi.org/10.3389/fmicb.2015.00235 PMID: 25914674

33. Chen H, Li F, Wang L, Jin Y, Chi C-H, Kurgan L, et al. Systematic evaluation of machine learning methods for identifying human–pathogen protein–protein interactions. Brief Bioinform. 2021; 22(3): bbaa068. https://doi.org/10.1093/bib/bbaa068 PMID: 32459334

34. Halder AK, Dutta P, Kundu M, Basu S, Nasipuri M. Review of computational methods for virus–host protein interaction prediction: a case study on novel Ebola–human interactions. Briefings in functional genomics. 2018; 17(6): 381–391. https://doi.org/10.1093/bfgp/elx026 PMID: 29028879

35. Cuevas JM, Geller R, Garijo R, López-Aldeguer J, Sanjuán R. Extremely high mutation rate of HIV-1 in vivo. PLoS Biol. 2015; 13(9): e1002251. https://doi.org/10.1371/journal.pbio.1002251 PMID: 26375597

36. Gao G, Wu X, Zhou J, He M, He JJ, Guo D. Inhibition of HIV-1 transcription and replication by a newly identified cyclin T1 splice variant. J Biol Chem. 2013; 288(20): 14297–14309. https://doi.org/10.1074/jbc.M112.438465 PMID: 23569210

37. Okada H, Zhang X, Fofana IB, Nagai M, Suzuki H, Ohashi T, et al. Synergistic effect of human CycT1 and CRM1 on HIV-1 propagation in rat T cells and macrophages. Retrovirology. 2009; 6(1): 1–12. https://doi.org/10.1186/1742-4690-6-43 PMID: 19435492

38. Kwon Y, Kaake RM, Echeverria I, Suarez M, Shamsabadi MK, Stoneham C, et al. Structural basis of CD4 downregulation by HIV-1 Nef. Nat Struct Mol Biol. 2020; 27(9): 822–828. https://doi.org/10.1038/s41594-020-0463-z PMID: 32719457

39. Jette CA, Barnes CO, Kirk SM, Melillo B, Smith AB, Bjorkman PJ. Cryo-EM structures of HIV-1 trimer bound to CD4-mimetics BNM-III-170 and M48U1 adopt a CD4-bound open conformation. Nat Commun. 2021; 12(1): 1–10. https://doi.org/10.1038/s41467-020-20314-w PMID: 33397941

40. Singha S, Shenoy PP. An adaptive heuristic for feature selection based on complementarity. Machine Learning. 2018; 107(12): 2027–2071. https://doi.org/10.1007/s10994-018-5728-y

41. Yeom S, Giacomelli I, Fredrikson M, Jha S, editors. Privacy risk in machine learning: Analyzing the connection to overfitting. 2018 IEEE 31st Computer Security Foundations Symposium (CSF); 2018: IEEE.

42. Ying X, editor An overview of overfitting and its solutions. Journal of Physics: Conference Series; 2019: IOP Publishing.

43. Lever J, Krzywinski M, Altman N. Points of significance: model selection and overfitting. Nature Publishing Group; 2016.

44. Ako-Adjei D, Fu W, Wallin C, Katz KS, Song G, Darji D, et al. HIV-1, human interaction database: current status and new features. Nucleic Acids Res. 2015; 43(D1): D566–D570. https://doi.org/10.1093/nar/gku1126 PMID: 25378338

45. Braschi B, Denny P, Gray K, Jones T, Seal R, Tweedie S, et al. Genenames. org: the HGNC and VGNC resources in 2019. Nucleic Acids Res. 2019; 47(D1): D786–D792. https://doi.org/10.1093/nar/gky930 PMID: 30304474

46. O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, et al. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res. 2016; 44(D1): D733–D745. https://doi.org/10.1093/nar/gkv1189 PMID: 26553804

47. Yates AD, Achuthan P, Akanni W, Allen J, Allen J, Alvarez-Jarreta J, et al. Ensembl 2020. Nucleic Acids Res. 2020; 48(D1): D682–D688. https://doi.org/10.1093/nar/gkz966 PMID: 31691826

48. Palasca O, Santos A, Stolte C, Gorodkin J, Jensen LJ. TISSUES 2.0: an integrative web resource on mammalian tissue expression. Database. 2018; 2018. https://doi.org/10.1093/database/bay028 PMID: 30403794

49. Consortium GO. The gene ontology resource: 20 years and still GOing strong. Nucleic Acids Res. 2019; 47(D1): D330–D338. https://doi.org/10.1093/nar/gky1055 PMID: 30395331

50. Alanis-Lobato G, Andrade-Navarro MA, Schaefer MH. HIPPIE v2.0: enhancing meaningfulness and reliability of protein–protein interaction networks. Nucleic Acids Res. 2016: gkw985. https://doi.org/10.1093/nar/gkw985 PMID: 27794551

51. Chang C-C, Lin C-J. LIBSVM: A library for support vector machines. ACM Trans Intell Syst Technol. 2011; 2(3): 1–27. https://doi.org/10.1145/1961189.1961199

52. Fu L, Niu B, Zhu Z, Wu S, Li W. CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics. 2012; 28(23): 3150–3152. https://doi.org/10.1093/bioinformatics/bts565 PMID: 23060610

53. Hakes L, Pinney JW, Lovell SC, Oliver SG, Robertson DL. All duplicates are not equal: the difference between small-scale and genome duplication. Genome Biol. 2007; 8(10): 1–13. https://doi.org/10.1186/gb-2007-8-10-r209 PMID: 17916239

54. King CR, Mehle A. The later stages of viral infection: An undiscovered country of host dependency factors. PLoS Pathog. 2020; 16(8): e1008777. https://doi.org/10.1371/journal.ppat.1008777 PMID: 32841303

55. Martinelli E, Cicala C, Van Ryk D, Goode DJ, Macleod K, Arthos J, et al. HIV-1 gp120 inhibits TLR9-mediated activation and IFN-α secretion in plasmacytoid dendritic cells. Proceedings of the National Academy of Sciences. 2007; 104(9): 3396–3401. https://doi.org/10.1073/pnas.0611353104 PMID: 17360657

56. Taylor HE, Khatua AK, Popik W. The innate immune factor apolipoprotein L1 restricts HIV-1 infection. J Virol. 2014; 88(1): 592–603. https://doi.org/10.1128/JVI.02828-13 PMID: 24173214

57. Kaul M, Ma Q, Medders K, Desai M, Lipton S. HIV-1 coreceptors CCR5 and CXCR4 both mediate neuronal cell death but CCR5 paradoxically can also contribute to protection. Cell Death Differ. 2007; 14 (2): 296–305. https://doi.org/10.1038/sj.cdd.4402006 PMID: 16841089

58. Liu S, Wang Q, Yu X, Li Y, Guo Y, Liu Z, et al. HIV-1 inhibition in cells with CXCR4 mutant genome created by CRISPR-Cas9 and piggyBac recombinant technologies. Sci Rep. 2018; 8(1): 1–11. https://doi.org/10.1038/s41598-017-17765-5 PMID: 29311619

59. Rodriguez JM, Rodriguez-Rivas J, Di Domenico T, Vázquez J, Valencia A, Tress ML. APPRIS 2017: principal isoforms for multiple gene sets. Nucleic Acids Res. 2018; 46(D1): D213–D217. https://doi.org/10.1093/nar/gkx997 PMID: 29069475

60. Jassal B, Matthews L, Viteri G, Gong C, Lorente P, Fabregat A, et al. The reactome pathway knowledgebase. Nucleic Acids Res. 2020; 48(D1): D498–D503. https://doi.org/10.1093/nar/gkz1031 PMID: 31691815

61. Gordon DE, Watson A, Roguev A, Zheng S, Jang GM, Kane J, et al. A quantitative genetic interaction map of HIV infection. Mol Cell. 2020; 78(2): 197–209. https://doi.org/10.1016/j.molcel.2020.02.004 PMID: 32084337

62. Wang ET, Sandberg R, Luo S, Khrebtukova I, Zhang L, Mayr C, et al. Alternative isoform regulation in human tissue transcriptomes. Nature. 2008; 456(7221): 470–476. https://doi.org/10.1038/nature07509 PMID: 18978772

63. Keren H, Lev-Maor G, Ast G. Alternative splicing and evolution: diversification, exon definition and function. Nature Reviews Genetics. 2010; 11(5): 345–355. https://doi.org/10.1038/nrg2776 PMID: 20376054

64. Takata MA, Gonçalves-Carneiro D, Zang TM, Soll SJ, York A, Blanco-Melo D, et al. CG dinucleotide suppression enables antiviral defence targeting non-self RNA. Nature. 2017; 550(7674): 124–127. https://doi.org/10.1038/nature24039 PMID: 28953888

65. Yu C-H, Dang Y, Zhou Z, Wu C, Zhao F, Sachs MS, et al. Codon usage influences the local rate of translation elongation to regulate co-translational protein folding. Mol Cell. 2015; 59(5): 744–754. https://doi.org/10.1016/j.molcel.2015.07.018 PMID: 26321254

66. Guéguen L, Duret L. Unbiased estimate of synonymous and nonsynonymous substitution rates with nonstationary base composition. Mol Biol Evol. 2018; 35(3): 734–742. https://doi.org/10.1093/molbev/msx308 PMID: 29220511

67. Betts MJ, Russell RB. Amino acid properties and consequences of substitutions. Bioinformatics for geneticists. 2003; 317: 289. https://doi.org/10.1002/0470867302.ch14

68. Vens C, Rosso M-N, Danchin EG. Identifying discriminative classification-based motifs in biological sequences. Bioinformatics. 2011; 27(9): 1231–1238. https://doi.org/10.1093/bioinformatics/btr110 PMID: 21372086

69. Dinkel H, Van Roey K, Michael S, Kumar M, Uyar B, Altenberg B, et al. ELM 2016—data update and new functionality of the eukaryotic linear motif resource. Nucleic Acids Res. 2016; 44(D1): D294–D300. https://doi.org/10.1093/nar/gkv1291 PMID: 26615199

70. Liu X-Y, Wu J, Zhou Z-H. Exploratory undersampling for class-imbalance learning. IEEE Trans Syst Man Cybern. 2008; 39(2): 539–550. https://doi.org/10.1109/TSMCB.2008.2007853 PMID: 19095540

71. Walsh I, Martin AJ, Di Domenico T, Tosatto SC. ESpritz: accurate and fast prediction of protein disorder. Bioinformatics. 2012; 28(4): 503–509. https://doi.org/10.1093/bioinformatics/btr682 PMID: 22190692

72. Mészáros B, Erdős G, Dosztányi Z. IUPred2A: context-dependent prediction of protein disorder as a function of redox state and protein binding. Nucleic Acids Res. 2018; 46(W1): W329–W337. https://doi.org/10.1093/nar/gky384 PMID: 29860432

73. Xue B, Mizianty MJ, Kurgan L, Uversky VN. Protein intrinsic disorder as a flexible armor and a weapon of HIV-1. Cell Mol Life Sci. 2012; 69(8): 1211–1259. https://doi.org/10.1007/s00018-011-0859-3 PMID: 22033837

74. King DF, Siddiqui AA, Buffa V, Fischetti L, Gao Y, Stieh D, et al. Mucosal tissue tropism and dissemination of HIV-1 subtype B acute envelope-expressing chimeric virus. J Virol. 2013; 87(2): 890–899. https://doi.org/10.1128/JVI.02216-12 PMID: 23135721

75. Bet A, Maze EA, Bansal A, Sterrett S, Gross A, Graff-Dubois S, et al. The HIV-1 antisense protein (ASP) induces CD8 T cell responses during chronic infection. Retrovirology. 2015; 12(1): 1–13. https://doi.org/10.1186/s12977-015-0135-y PMID: 25809376

76. Ahmed H, Howton T, Sun Y, Weinberger N, Belkhadir Y, Mukhtar MS. Network biology discovers pathogen contact points in host protein-protein interactomes. Nat Commun. 2018; 9(1): 1–13. https://doi.org/10.1038/s41467-017-02088-w PMID: 29317637

77. Cafarelli T, Desbuleux A, Wang Y, Choi SG, De Ridder D, Vidal M. Mapping, modeling, and characterization of protein–protein interactions on a proteomic scale. Curr Opin Struct Biol. 2017; 44: 201–210. https://doi.org/10.1016/j.sbi.2017.05.003 PMID: 28575754

78. Doncheva NT, Assenov Y, Domingues FS, Albrecht M. Topological analysis and interactive visualization of biological networks and protein structures. Nat Protoc. 2012; 7(4): 670. https://doi.org/10.1038/nprot.2012.004 PMID: 22422314

79. Liu Z. A method of SVM with normalization in intrusion detection. Procedia Environmental Sciences. 2011; 11: 256–262. https://doi.org/10.1016/j.proenv.2011.12.040

80. Babbar R, Schölkopf B. Data scarcity, robustness and extreme multi-label classification. Machine Learning. 2019; 108(8): 1329–1351. https://doi.org/10.1007/s10994-019-05791-5

81. Cheng Q, Zhou H, Cheng J. The fisher-markov selector: fast selecting maximally separable feature subset for multiclass classification with applications to high-dimensional data. IEEE Trans Pattern Anal Mach Intell. 2010; 33(6): 1217–1233. https://doi.org/10.1109/TPAMI.2010.195 PMID: 21493968

82. Zhang J, Chai H, Gao B, Yang G, Ma Z. HEMEsPred: Structure-based ligand-specific heme binding residues prediction by using fast-adaptive ensemble learning scheme. IEEE/ACM Trans Comput Biol Bioinform. 2016; 15(1): 147–156. https://doi.org/10.1109/TCBB.2016.2615010 PMID: 28029626

83. Chai H, Zhang J. Identification of mammalian enzymatic proteins based on sequence-derived features and species-specific scheme. IEEE Access. 2018; 6: 8452–8458. https://doi.org/10.1109/ACCESS.2018.2798284

84. Chicco D, Jurman G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. BMC Genomics. 2020; 21(1): 1–13. https://doi.org/10.1186/s12864-019-6413-7 PMID: 31898477

85. Gautier VW, Gu L, O'Donoghue N, Pennington S, Sheehy N, Hall WW. In vitro nuclear interactome of the HIV-1 Tat protein. Retrovirology. 2009; 6(1): 1–18. https://doi.org/10.1186/1742-4690-6-47 PMID: 19454010

86. Fang M, Xu N, Shao X, Yang J, Wu N, Yao H. Inhibitory effects of human immunodeficiency virus gp120 and Tat on CpG-A-induced inflammatory cytokines in plasmacytoid dendritic cells. Acta Biochim Biophys Sin. 2012; 44(9): 797–804. https://doi.org/10.1093/abbs/gms062 PMID: 22814248

87. Barr SD, Smiley JR, Bushman FD. The interferon response inhibits HIV particle production by induction of TRIM22. PLoS Pathog. 2008; 4(2): e1000007. https://doi.org/10.1371/journal.ppat.1000007 PMID: 18389079

88. Gao D, Wu J, Wu Y-T, Du F, Aroh C, Yan N, et al. Cyclic GMP-AMP synthase is an innate immune sensor of HIV and other retroviruses. Science. 2013; 341(6148): 903–906. https://doi.org/10.1126/science.1240933 PMID: 23929945

89. Mangino G, Percario ZA, Fiorucci G, Vaccari G, Manrique S, Romeo G, et al. In vitro treatment of human monocytes/macrophages with myristoylated recombinant Nef of human immunodeficiency virus type 1 leads to the activation of mitogen-activated protein kinases, IκB kinases, and interferon regulatory factor 3 and to the release of beta interferon. J Virol. 2007; 81(6): 2777–2791. https://doi.org/10.1128/JVI.01640-06 PMID: 17182689

90. Yim HC, Li JC, Lau JS, Lau AS. HIV-1 Tat dysregulation of lipopolysaccharide-induced cytokine responses: microbial interactions in HIV infection. AIDS. 2009; 23(12): 1473–1484. https://doi.org/10.1097/QAD.0b013e32832d7abe PMID: 19622906

91. Harman AN, Nasr N, Feetham A, Galoyan A, Alshehri AA, Rambukwelle D, et al. HIV blocks interferon induction in human dendritic cells and macrophages by dysregulation of TBK1. J Virol. 2015; 89(13): 6575–6584. https://doi.org/10.1128/JVI.00889-15 PMID: 25855743

92. Bego MG, Côté É, Aschman N, Mercier J, Weissenhorn W, Cohen ÉA. Vpu exploits the cross-talk between BST2 and the ILT7 receptor to suppress anti-HIV-1 responses by plasmacytoid dendritic cells. PLoS Pathog. 2015; 11(7): e1005024. https://doi.org/10.1371/journal.ppat.1005024 PMID: 26172439

93. Enard D, Cai L, Gwennap C, Petrov DA. Viruses are a dominant driver of protein adaptation in mammals. Elife. 2016; 5: e12469. https://doi.org/10.7554/eLife.12469 PMID: 27187613

94. Pearson WR. Finding protein and nucleotide similarities with FASTA. Curr Protoc Bioinformatics. 2016; 53(1): 3–9. https://doi.org/10.1002/0471250953.bi0309s53 PMID: 27010337

95. Maetschke SR, Simonsen M, Davis MJ, Ragan MA. Gene Ontology-driven inference of protein–protein interactions using inducers. Bioinformatics. 2012; 28(1): 69–75. https://doi.org/10.1093/bioinformatics/btr610 PMID: 22057159

96. Puntel M, Barrett R, Sanderson NS, Kroeger KM, Bondale N, Wibowo M, et al. Identification and visualization of CD8+ T cell mediated IFN-γ signaling in target cells during an antiviral immune response in the brain. PLoS One. 2011; 6(8): e23523. https://doi.org/10.1371/journal.pone.0023523 PMID: 21897844

97. Caby F. CD4+/CD8+ ratio restoration in long-term treated HIV-1-infected individuals. AIDS. 2017; 31(12): 1685–1695. https://doi.org/10.1097/QAD.0000000000001533 PMID: 28700392

98. Mohanapriya M, Lekha J, editors. Comparative study between decision tree and knn of data mining classification technique. Journal of Physics: Conference Series; 2018: IOP Publishing.

99. Han S, Kim H, Lee Y-S. Double random forest. Machine Learning. 2020; 109(8): 1569–1586. https://doi.org/10.1007/s10994-020-05889-1

100. Schnoes AM, Brown SD, Dodevski I, Babbitt PC. Annotation error in public databases: misannotation of molecular function in enzyme superfamilies. PLoS Comput Biol. 2009; 5(12): e1000605. https://doi.org/10.1371/journal.pcbi.1000605 PMID: 20011109

101. Churchill MJ, Deeks SG, Margolis DM, Siliciano RF, Swanstrom R. HIV reservoirs: what, where and how to target them. Nat Rev Microbiol. 2016; 14(1): 55–60. https://doi.org/10.1038/nrmicro.2015.5 PMID: 26616417

102. Harlalka GV, Baple EL, Cross H, Kühnle S, Cubillos-Rojas M, Matentzoglu K, et al. Mutation of HERC2 causes developmental delay with Angelman-like features. J Med Genet. 2013; 50(2): 65–73. https://doi.org/10.1136/jmedgenet-2012-101367 PMID: 23243086

103. Sathasivam K, Neueder A, Gipson TA, Landles C, Benjamin AC, Bondulich MK, et al. Aberrant splicing of HTT generates the pathogenic exon 1 protein in Huntington disease. Proceedings of the National

Academy of Sciences. 2013; 110(6): 2366–2370. https://doi.org/10.1073/pnas.1221891110 PMID: 23341618

104. Rose M, Schubert C, Dierichs L, Gaisa NT, Heer M, Heidenreich A, et al. OASIS/CREB3L1 is epigenetically silenced in human bladder cancer facilitating tumor cell spreading and migration in vitro. Epigenetics. 2014; 9(12): 1626–1640. https://doi.org/10.4161/15592294.2014.988052 PMID: 25625847

105. Khan HA, Margulies CE. The role of mammalian Creb3-like transcription factors in response to nutrients. Front Genet. 2019; 10: 591. https://doi.org/10.3389/fgene.2019.00591 PMID: 31293620

106. Qiu J, Liang T, Wu J, Yu F, He X, Tian Y, et al. N-Substituted Pyrrole Derivative 12m Inhibits HIV-1 Entry by Targeting Gp41 of HIV-1 Envelope Glycoprotein. Front Pharmacol. 2019; 10. https://doi.org/10.3389/fphar.2019.00859 PMID: 31427969

107. Jordan IK, Wolf YI, Koonin EV. Duplicated genes evolve slower than singletons despite the initial rate increase. BMC Evol Biol. 2004; 4(1): 1–11. https://doi.org/10.1186/1471-2148-4-22 PMID: 15238160

108. Wibmer CK, Gorman J, Ozorowski G, Bhiman JN, Sheward DJ, Elliott DH, et al. Structure and recognition of a novel HIV-1 gp120-gp41 interface antibody that caused MPER exposure through viral escape. PLoS Pathog. 2017; 13(1): e1006074. https://doi.org/10.1371/journal.ppat.1006074 PMID: 28076415

109. Kanehisa M, Furumichi M, Tanabe M, Sato Y, Morishima K. KEGG: new perspectives on genomes, pathways, diseases and drugs. Nucleic Acids Res. 2017; 45(D1): D353–D361. https://doi.org/10.1093/nar/gkw1092 PMID: 27899662