Ensemble Learning: Predicting Human Pathogenicity of Hematophagous Arthropod Vector-Borne Viruses

Huakai Hu[1, 2, #], Chaoying Zhao[3, 1, #], Meiling Jin[4], Jiali Chen[5], Xiong Liu[1], Hua Shi[1], Jinpeng Guo[1], Changjun Wang[3, 1, *], Yong Chen[1, 2, *]

[1]School of Public Health, China Medical University, Shenyang, Liaoning province, China

[2]China Chinese PLA Center for Disease Control and Prevention, Beijing, China

[3]School of Public Health, Zhengzhou University, Zhengzhou, Henan Province, China

[4]Liaoning Provincial Center for Disease Control and Prevention, Shenyang, Liaoning, People's Republic of China

[5]School of Medicine, NanKai University, Tianjin, People's Republic of China.

Running title: Predicting Pathogenicity of Hematophagous Arthropod Vector-Borne Viruses

[*]These authors contributed equally to this work and are listed as co-first authors

Correspondence to Yong Chen chenyonger@126.com

Changjun Wang, science2008@hotmail.com

Abstract

Hematophagous arthropods occupy a pivotal role in ecosystems, serving as vectors for a wide array of pathogens with significant implications for public health. Their capacity to harbor and transmit viruses through biting actions creates a substantial risk of zoonotic spillover. Despite the advancements in metagenomic approaches for virus discovery in vectors, the isolation and cultivation of viruses still pose significant challenges, thereby limiting comprehensive assessments of their pathogenicity. Here, we curated two datasets: one with 294 viruses, characterized by 37 epidemiological features, encompassing virus information and host associations; the second with 71,622 sequences of hematophagous arthropod vector-borne viruses, annotated with 33 sequence features. Two XGBoost models were

developed to predict arbovirus human pathogenicity—one integrating macroscopic eco-epidemiological data, the other incorporating virus-related sequence features. The macroscopic model identified non-vector host transmission as a key determinant, especially involving Perissodactyla, Artiodactyla, and Carnivora Order. The sequence-based model demonstrated that viral adhesion and viral invasion had distinct trends with consistent increase and decrease in the likelihood of virus pathogenicity to humans, respectively. With validated through an independent dataset, the model exhibited a congruous alignment with documented pathogenicity outcomes. Together, the models offer a holistic framework for assessing the pathogenic potential of viruses transmitted by hematophagous arthropods.

Introduction

Hematophagous arthropods, such as mosquitoes and ticks, play a pivotal role in ecosystems as blood consumers and crucial disease vectors (Cuthbert et al., 2023; Touray et al., 2023). These arthropods can harbor a myriad of pathogens, including bacteria, fungi, and viruses. Notably, viral infections in these organisms are classified under the umbrella terms of Arthropod-Borne Viruses (arboviruses) and insect-specific viruses (ISVs) (Calisher & Higgs, 2018; Gould et al., 2017; Nouri et al., 2018; Zhao et al., 2022). The potential for these arthropods to harbor and disseminate a diverse array of pathogens poses a grave threat to both human and animal health, with the ominous potential to trigger outbreaks and result in a substantial number of annual fatalities (Batson et al., 2021; Roth et al., 2018). Vector-borne diseases contribute significantly to infectious diseases (Chala & Hamde, 2021), with notable arboviruses including the Zika virus (Khongwichit et al., 2023; Weaver et al., 2018), Japanese encephalitis virus (JEV) (Kampen & Werner, 2014), and the incessant menace of Dengue virus (DENV) (Fournet et al., 2023).

In recent years, propelled by the widespread adoption of Viral Metagenomics sequencing technologies, the identification of a wide range of established and emerging viruses within hematophagous vectors, such as mosquitoes and ticks, has become feasible (Ni et al., 2023; X. Yang et al., 2023). This technological progress presents an unprecedented opportunity to

comprehensively explore the distribution and transmission patterns of arboviruses and ISVs across a spectrum of hosts, including both vectors and non-vectors. Such advancements are crucial for supporting early warning systems, facilitating the anticipation and mitigation of disease spread before its onset (Birnberg et al., 2020; Brinkmann et al., 2016). Despite these achievements in viral metagenomics, current bioinformatic methods for virus recognition still face limitations (Fang et al., 2019). Accurate identification of a significant number of unknown contigs remains challenging. Even when identifying known or novel viruses, the direct isolation and cultivation of these viruses from vectors proves to be formidable tasks, hindering in-depth exploration of their pathogenesis and immune response (Lewis et al., 2021).

In general, the close phylogenetic relatedness among viruses can offer insights into their potential for human infectivity, as closely related viruses are generally presumed to share common phenotypes and host ranges (Geoghegan & Holmes, 2018). However, despite being a common rule of thumb for virus risk assessment, the extent to which evolutionary proximity to viruses with known human infectivity accurately predicts zoonotic potential remains unexamined in the current literature (Behl et al., 2022). Furthermore, the specific model is designed to be trained on sequence features of closely related viruses (i.e., strains of the same species) to discern viruses with human infectivity (Zhang et al., 2019). Unfortunately, this method often overlooks critical functional characteristics of the viral genome, resulting in a model that is less inclined to identify universally applicable pathogenic features across diverse viruses. Consequently, predictions derived from such a model may be highly susceptible to substantial biases (Mollentze et al., 2021).

The epidemiological characteristics of virus transmission encompass not only the virus itself and information about its vector host (Zaid et al., 2021; Y.-J. S. Huang et al., 2019a; Viglietta et al., 2021) but also factors such as geographical and climatic variations, as well as interactions with non-vector hosts (Ciota & Keyel, 2019; Conway et al., 2014; Forrester et al., 2014; Tabachnick, 2016). Moreover, for specific viruses, their nucleotide sequence information may reflect actual pathogenic details (Bartoszewicz, Genske, et al., 2021). Therefore, through a comprehensive analysis that integrates both macroscopic and microscopic perspectives, our objective is to identify the epidemiological features and viral

sequence characteristics that have the greatest impact on the potential pathogenicity to humans.

Based on the global data of arthropod-borne virus compiled by Huang et al. (Y. Huang et al., 2023) as a foundation, we carefully curated the contents to extract pertinent information concerning hematophagous arthropod-borne viruses. Additionally, to augment our analysis, we utilized SeqScreen for insightful functional details of the viral sequences (Balaji et al., 2022). Employing the XGBoost algorithm with ensemble learning, we developed both regression and classification prediction models. This facilitated the identification of factors with the most significant impact on human pathogenicity and enabled the construction of ensemble learning for predicting the pathogenicity of virus sequences carried by hematophagous arthropods.

Materials and methods

Database restructuring and epidemiological feature retrieval

The initial dataset comprised 101,094 virus sequences sourced from NCBI, spanning the period from March 11, 1991, to January 28, 2023 (Y. Huang et al., 2023). To enhance the reliability and specificity of our analysis, a stringent screening process was applied, systematically excluding records lacking host information, sampling location details, and those with ambiguous vector-host relationships. It is important to note that this dataset excludes data from Antarctica. Subsequently, we identified 11 species of hematophagous arthropods, including mosquitoes and ticks (Table supplement 1), while excluding non-blood-feeding species such as *Tipulidae* and *Chironomidae*. Following this, we systematically screened the entire dataset, retaining records exclusively related to hosts classified as hematophagous arthropods. This refined dataset, derived through meticulous curation, forms the foundation for our research, ensuring the integrity and accuracy of subsequent analyses. Nevertheless, due to inaccuracies in the classification of vectors within the database, a Python script was developed to scrape taxonomic directory. This script retrieved detailed order, family, and genus information for each hematophagous arthropod and non-vector host classification. To reveal the distinct composition of non-vector hosts, host

counts underwent logarithmic transformation (Figure 1A). For a more comprehensive presentation, host classifications with fewer than 100 occurrences were amalgamated into an "others" category, resulting in a total of 10 host classifications (Figure 1C).

In terms of additional epidemiological features, Köppen climate classification data for each vector were acquired based on their discovery locations. This information was sourced from both the Weather and Climate website (https://weatherandclimate.com/) and the Mindat website (https://www.mindat.org/). Concurrently, continental data for each country were obtained from the World Population Review (https://worldpopulationreview.com/continents), and Baltimore classification data were sourced from the International Committee on Taxonomy of Viruses (ICTV) (https://ictv.global/report/genome).

The development of a regression model for macroscopic characteristics

Firstly, among the 8,468 datasets in this study, We employed an R script to transform it into a dataset comprising 294 distinct virus types, each characterized by 37 unique features (Table 1). Subsequently, we utilized the XGBoost ensemble learning model to establish regression models. The dataset was divided into training and validation sets at a ratio of 7.5:2.5. Given the balanced ratio of positive to negative samples (1:1) in the model's database, addressing imbalance was not deemed necessary. The training set was employed to train the model based on the specified parameters (Table supplement 2). After determining the optimal number of iterations through 10-fold cross-validation, we proceeded to construct the final model using this identified count.

Development and validation of a macroscopic features classification model

In the microscopic pathogenicity classification model, we annotated the aforementioned database uniformly using SeqScreen, resulting in a total of 71,622 virus sequences. After excluding viruses from hosts submitted to NCBI after 2022, we obtained a final dataset of 71,593 sequences for this model. Due to the imbalance in positive and negative samples in the database (positivity rate of 79.3%), we adjusted the sample sampling rate to balance the dataset. The specific parameters employed in this model are meticulously detailed in Table supplement 3. Utilizing the training set, the model was trained in accordance with these parameters, determining the optimal iteration count through rigorous 5-fold cross-validation. Subsequently, the final model was constructed utilizing the identified optimal iteration count.

To constitute an additional validation dataset, we retrieved Ebinur Lake Virus with arthropods as hosts from NCBI, incorporating these samples with those previously excluded. Following a consistent application of the specified parameters, we trained the model using functional features from the entire dataset. Subsequently, predictions were generated on the additional validation dataset. The obtained results underwent a meticulous comparative analysis with findings from established pathogenic studies.

Results

Global overview of hematophagous vector-virus distribution, diversity, and host interactions

This study has curated a comprehensive dataset of 8,468 hematophagous vector-virus pairs, shedding light on their geographical distribution, diversity, and interactions with hosts. In terms of distribution, these vectors were classified into two principal classes: *Insecta* and *Arachnida*, spanning seven distinct families (Figure 1A). The records cover all six continents except Antarctica, spanning across 102 countries globally and representing 24 diverse climate types (Figure 1B). Regarding diversity, among the hematophagous vectors, *Culicidae* (64%, 5,445 sequences) predominates, constituting over half of all records, followed by *Ixodidae* (32%, 2,703 sequences). Globally, the United States exhibits the highest diversity and abundance of vectors, hosting five distinct families, followed by China with four. In terms of virus records associated with vectors, the United States (1,977) leads the list, followed by Russia, China, and Japan.

Turning to non-vector hosts, the dataset includes an additional 54,789 pairs of non-vector hosts and viruses, with the non-vector hosts categorized into 15 groups. Among these hosts, humans are the most prevalent, accounting for 40,078 records, followed by *Artiodactyla* and *Aves*, constituting nearly 20% of the total (Figure 1C). The interactions between viruses and non-vector hosts are distinct. The majority of viruses are associated with a single host. Notably, West Nile virus (WNV) and Tick-borne encephalitis virus (TBEV), both belonging to the *Flaviviridae*, exhibit the most widespread cross-host transmission, being detected in

nine non-vector host species. Moreover, as viruses expand their capacity to infect a wide range of non-vector hosts, a noticeable reduction in viral diversity is observed. Specifically, viruses capable of infecting only one host encompass 10 distinct virus families, while those exhibiting infectivity across two to four hosts are confined to five families. Remarkably, viruses with the ability to infect five, six, or seven hosts are prominently represented by families such as *Flaviviridae* and *Togaviridae*. Among viruses capable of infecting seven hosts, Dabie bandavirus stands out as a unique case. Belonging to the *Phenuiviridae*, this virus is predominantly found in Asia (China, Japan, and South Korea). Infection with Dabie bandavirus poses a significant health risk, causing a severe febrile illness accompanied by thrombocytopenia, known as Severe Fever with Thrombocytopenia Syndrome (SFTS), leading to its alternate nomenclature as the SFTS virus. The Japanese encephalitis virus exhibits the highest degree of cross-vector host diversity, being detectable in three distinct vector families: *Culicidae*, *Ixodidae*, and *Ceratopogonidae*. The previously mentioned WNV and TBEV demonstrate transmission capabilities across both vector families, *Culicidae* and *Ixodidae*.

Pathogenicity of hematophagous arthropod vector-borne viruses: a macroscopic regression analysis of epidemiological characteristics

Through transforming the mentioned database and enhancing it with additional epidemiological characteristics, we constructed a comprehensive dataset for the model. This dataset comprises 294 distinct viruses, each characterized by 37 diverse features, broadly categorized into viral characteristics, vector host features, and non-vector host features. To unpack the crucial factors underlying human pathogenicity, we constructed and rigorously trained an XGBoost model. This model leverages human infection status as the dependent variable and incorporates 36 diverse features as independent variables, pinpointing the key determinants of human infection. The model exhibits robust performance on the testing set, with minimal prediction errors reflected in low MSE (0.01) and MAE (0.05) values, highlighting its accuracy. Additionally, high R^2 (94.20%) and Explained Variance (94.29%) values underscore the model's comprehensive ability to explain the variance in the dependent

variable.

The detection of viruses in non-vector hosts significantly influences human pathogenicity (Figure 2), surpassing the impact of both vector hosts and the viral agents themselves. Notably, the characteristics “Cross_host”, representing the total diversity of non-vector hosts in which the virus has been detected, carries a weight of 52 in the model. This underscores the critical role of the diversity of non-vector hosts in determining human pathogenicity. Specifically, when considering potential human-pathogenicity, the order of importance is as follows: *Perissodactyla*, *Artiodactyla Carnivora* and *Aves*. The higher the diversity of virus detections in these animals, the greater the likelihood of the virus being pathogenic to humans. After non-vector host factors, the subsequent important set of characteristics relates to the vector hosts. Among these, “Cross_vector_g”, representing interspecies transmission among diverse vector genus, emerges as the most critical factor. If a virus can propagate within diverse vector genus, there is a substantial likelihood of viral spillover. The third set of characteristics relates to the intrinsic characteristics of the virus itself. The closer the viral phylogenetic relationship, the higher the likelihood of inducing similar immune responses, thereby leading to diseases.

Relationship between viral genomic function and human pathogenicity: a microscopic machine learning approach

In our research, we employed SeqScreen to functionally annotate all viral sequences in our comprehensive dataset. After excluding sequences without successful annotations, our refined dataset comprised 71,622 arboviruses and ISV sequences, each accompanied by their respective functional, host, and pathogenic features. The largest category within our dataset consists of mosquito-borne arboviruses, with Dengue virus 1 (9,194 sequences), Dengue virus 2 (8,999 sequences), and West Nile virus (4,656 sequences) being the most prevalent. Tick-borne arboviruses, including African swine fever virus (3,915 sequences) and Crimean-Congo hemorrhagic fever orthonairovirus (3,771 sequences) closely follow in quantity.

Our functional annotation revealed a total of 10 distinct pathogenic features. The results indicated that “viral adhesion” is the most prevalent function, accounting for 62% (44,482

sequences). This function facilitates virus adhesion to host cells, initiating infection and paving the way for subsequent invasion and replication. Following closely are the “viral counter-signaling” (49%) and “host xenophagy” (47%), which are typically associated with immune evasion. These mechanisms enable the virus to survive, replicate within host cells, and successfully transmit to other cells (Table 1).

Among known non-pathogenic viruses to humans, “viral invasion” stands out as the most prevalent function, despite its relatively lower overall count compared to other functions. Notably, within these viruses, the primary hosts targeted are hematophagous arthropod vectors, with non-vector hosts predominantly represented by *Artiodactyla* and *Aves* (Figure 3A). Conversely, within the known human-pathogenic viruses, “viral adhesion” ranks as the most prevalent function in terms of annotation quantities. In this context, excluding human hosts, hematophagous arthropod vectors continue to be predominant, followed by *Aves* (Figure 3B). This observation suggests a potential genomic similarity in the pathogenicity of arboviruses among humans, hematophagous arthropod vectors and *Aves*.

We developed a binary classification XGBoost model using 33 features, which included functional annotations for all viruses in the database and viral size (length of virus). The model's dependent variable denotes whether a virus is pathogenic to humans. After excluding the viruses in the extra validation dataset, the remaining viruses in the database were allocated to a training set and a testing set in a 7.5:2.5 ratio. While achieving a high accuracy (95.36%), we incorporated additional metrics, such as Precision (97.57%), Recall (96.55%), and F1 score (97.06%), for a more nuanced assessment. Furthermore, we generated an ROC curve (Figure 4A) and a confusion matrix (Figure 4B) to gain a holistic view of the model's strengths and weaknesses.

The model’s results clearly demonstrate that, in terms of average gain, “viral adhesion” exhibits the highest value, significantly enhancing the model's predictive accuracy. “Host xenophagy” and “viral invasion” closely follow suit (Figure 5A). Regarding the model's coverage, “viral invasion” and “host ubiquitin” occupy the top two positions due to their capability to impact a wide array of viral sequences (Figure 5B). In terms of the model's weights, the size of the viral sequence takes precedence over other features, indicating its frequent utilization in the model construction and its vital role in making supplementary

assessments on the virus pathogenicity based on functional insights (Figure 5C).

In our conclusive analysis, we employed SHAP (SHapley Additive exPlanations) to gain deeper insights into the individual feature contributions to the model. Notably, the top-ranking feature— “viral size” —does not exhibit a discernible trend in pathogenicity to humans. However, other features reveal intriguing patterns. Specifically, both “viral adhesion” and “host xenophagy”, although slightly less significant than size, individually demonstrate distinct trends: viral sequences annotated with either of these functions consistently increase the likelihood of virus pathogenicity in humans. Conversely, “viral invasion” demonstrates an inverse relationship, wherein sequences possessing this trait tend to reduce the probability of virus pathogenicity. The majority of remaining features, on the other hand, positively correlate with pathogenicity. In summary, most features contribute towards determining the likelihood of virus pathogenicity in humans (Figure 6).

To delve deeper into the intricate interactions among these features, we conducted a thorough analysis. Our results highlight that among all features, “viral counter signaling” exhibits the most significant interaction with viral size. However, its impact on pathogenicity remains inconclusive, lacking a definitive directional trend (Figure 7A). Additionally, we observed a noteworthy interaction between “host xenophagy” and “viral adhesion”. The concurrent presence of these features substantially enhances the virus's pathogenicity towards humans (Figure 7B). Interestingly, “viral invasion” demonstrates a strong but contrasting interaction with “viral counter signaling”. Specifically, “viral counter signaling” seems to function as a protective factor against human pathogenicity when “viral invasion” is present, leading to a reduced likelihood of the virus being pathogenic to humans (Figure 7C).

In our comprehensive analysis of interactions across all features, significant insights emerged. While viral size lacks a clear discernible trend in its interaction with other features, the interplay of 'host xenophagy' with both “viral adhesion” and 'viral counter signaling' guides the model toward non-pathogenicity predictions, acting as a protective feature (Figure S1).

To assess real-world performance, we compiled an additional dataset consisting of 29 viruses carried by hematophagous vectors, submitted after 2022. This dataset comprises 24 strains of SFTS virus, 3 of Restan viruses, 3 of Tataguine viruses, 1 of Japanese encephalitis virus (JEV), and 1 of Nairobi sheep disease virus (NSDV). Furthermore, 25 sequences of Ebinur

Lake virus (EBIV) borne by arthropods were downloaded from the NCBI, resulting in a total of 54 virus sequences. To ensure independence, the content of the additional dataset was excluded from the original model database. The train and test datasets were merged to train the ultimate model. Validation was then conducted using the additional dataset. Model predictions indicated that all sequences of SFTSV, a specific strain of JEV, a particular variant of Tataguine virus, and one isolate of Ebinur Lake Virus potentially exhibit pathogenicity to humans.

Discussion

The intricate interplay between hematophagous arthropods and the viruses they harbor forms a dynamic ecosystem with profound implications for public health. Ticks and mosquitoes, integral components of ecosystems, serve as potent disease vectors capable of transmitting various pathogens, including arboviruses and ISVs. Recent advancements in Viral Metagenomics sequencing technologies have significantly transformed our investigation of the virome within hematophagous vectors, providing unparalleled access to comprehensive viral genetic information. Although technological advancements have been made, bioinformatic methods for virus recognition still face inherent limitations, particularly in identifying unknown contigs, which hinders comprehensive virome characterization. The isolation and cultivation of known or novel viruses from vectors remains challenging, impeding in-depth exploration of their pathogenesis and immune response. The common approach to assessing pathogenicity relies on viral phylogenetic analysis, assuming that viruses with significant phylogenetic distance share similar pathogenic properties. However, the extent to which phylogenetic relatedness accurately predicts the potential for zoonotic diseases remains a critical aspect requiring deeper exploration. Existing models, often tailored for closely related viruses, may inadvertently overshadow crucial functional characteristics, introducing biases in predictions. To address these challenges, this study innovatively adopts an ensemble learning algorithm in machine learning. Our objective is to comprehensively explore the human-pathogenicity of viruses borne by hematophagous vectors, considering

both macro and micro-level characteristics, with the ultimate aim of identifying key viral features associated with pathogenicity.

Based on the curated dataset, the distribution and abundance of hematophagous arthropods suggest that the United States, Russia, and China harbor the highest number of vector insects, notably mosquitoes and ticks, acting as primary carriers for arthropod-borne viruses (Y.-J. S. Huang et al., 2019b; Wu et al., 2023). Among the viruses carried by these arthropods, RNA viruses, specifically those classified as dsRNA under the third Baltimore group, dominate. *Phlebotominae* and *Ceratopogonidae* share virus families, while distinct mosquito genera harbor a diverse array of viruses, primarily belonging to the *Flaviviridae*, *Togaviridae*, and *Peribunyaviridae* (Figure supplement 1). The abundance of viruses is influenced by various factors, prompting a correlation analysis on the dataset. The results reveal strong correlations for most viruses, excluding *Asfarviridae*, Zirqa virus, and Wallerfield virus, with six key characteristics, including vector family and weather conditions. Within the community of vector-borne viruses, *Flaviviridae*, *Togaviridae*, *Bunyaviridae*, *Rhabdoviridae*, and *Phenuiviridae* are commonly co-detected and considered as core virome(Coatsworth et al., 2022) (Figure supplement 2).

Within the cyclic dynamics of arboviruses, numerous factors intricately influence the transmission to humans, impacting pathogenic outcomes. Variations in the intrinsic nature of viruses yield diverse levels of human pathogenicity, commonly associated with phylogenetic proximity. Moreover, an increased diversity of viruses within vectors may foster co-infection, thereby facilitating viral evolution and spill-over (Vogels et al., 2019). Consistent with these observations, our results highlight the significant influence of viral intrinsic factors on human pathogenicity. These viruses primarily propagate diseases through hematophagous vectors, predominantly mosquitoes, The species and behaviors of these vectors, collectively termed “Vector capacity” (Conway et al., 2014), along with environmental shifts (Hermanns et al., 2023; Weissenböck et al., 2010), play a pivotal role in shaping vector composition and consequentially impact viral transmission. Notably, our research reveals that the impact of vector hosts is equivalent to that of viral intrinsic factors. Both the diversity across vector genera and the quantity of vector genera exert a substantial influence on human pathogenicity.

Interactions between viruses and non-vector hosts drive viral evolution (Sen et al., 2016),

with interspecies interactions being the primary driver for viral spill-over (Y.-J. S. Huang et al., 2019a). The interplay between vectors and *Aves* hosts enables long-distance viral transmission (Forrester et al., 2014), while interactions with vertebrates also emerge as pivotal determinants(García-Romero et al., 2023; Golnar et al., 2014; Stephenson et al., 2019). Our results align with these insights, highlighting that, beyond interspecies categories, viruses infecting *Perissodactyla* and *Artiodactyla* pose the most significant risk for human pathogenicity, increasing the likelihood of transmission to humans and subsequent disease outbreaks.

Capitalizing on the ability to analyze viral genomic data for predicting viral pathogenicity to humans (Bartoszewicz, Seidel, et al., 2021), we conducted an exhaustive investigation into the genetic functionalities of arboviruses and ISVs within the database. Utilizing the ensemble learning algorithm, we meticulously developed and trained a predictive model. Furthermore, an additional dataset comprising arboviruses submitted to NCBI after 2022 was incorporated for validation and prediction purposes. Our model demonstrates superior performance, showcasing distinctive contributions from individual functional features that collectively shape the overarching trend in viral pathogenicity. Specifically, “viral adhesion”, representing a pivotal mechanism for viral infection and entry into host cells, emerges not only as the predominant feature but also significantly enhances the overall performance of the model. Empirical evidence affirms that the presence of this feature in viral sequences, subsequent to transmission to humans by hematophagous vectors, consistently indicates an elevated risk of pathogenicity. For instance, viruses within the *Flaviviridae*, such as DENV, WNV, and ZIKV (Begum et al., 2019; Cruz-Oliveira et al., 2015; Faustino et al., 2019; Hasan et al., 2017; Martins et al., 2019), utilize E and capsid proteins to enter receptor cells. Likewise, the Chikungunya virus, a member of the *Togaviridae*, facilitates the fusion with receptor cells through trimeric E1/E2 spikes (Ciota & Keyel, 2019).

While “viral invasion” plays a pivotal role in the initial phase of viral entry, it is notably scarce among this dataset (Table 1). Furthermore, most viruses with this feature are presently classified as non-pathogenic to humans (Figure 3). The high abundance of *Flavivirus* in the dataset could explain the limited occurrence of “viral invasion”, given their unique infection mechanisms that may not require this specific feature. Additionally, SeqScreen might not

have detected “viral invasion” in these sequences. Despite its relatively low occurrence in the dataset, this feature exhibits the highest cover value in the XGBoost model, indicating its significant impact on the model's performance. Interestingly, contrary to expectations, its presence predominantly acts as a “protective factor” in predicting pathogenicity, as revealed by SHAP explanations. This phenomenon may be attributed to the prevalence of this feature among non-pathogenic viruses in our dataset. However, since we balanced the samples during model training, it could also be a result of interactions among different features. A more in-depth exploration of interactions related to this trait revealed a robust interplay with “viral counter signaling”. When both features coexist, the model significantly leans towards predicting non-pathogenicity in humans. Importantly, these two processes are not mutually exclusive factors in actual virus infections. This observation implies potential distinctive invasion mechanisms of arboviruses, indicating unconventional pathways for entering host cells that facilitate immune evasion.

In this dataset, “viral counter signaling” and “host xenophagy” are prevalent features actively enhancing virus pathogenicity and triggering host infection. They play a crucial role in the pathogenicity to humans (Costa et al., 2013; King et al., 2020). Notably, “host xenophagy”, similar to “viral adhesion”, significantly influences the model results. In terms of interactions, it has the strongest interaction with “viral adhesion”, leading to a positive inclination towards pathogenicity.

The feature of “size”, representing the length of viral sequences, while not directly associated with pathogenic functions, plays a crucial role in refining the final results, as indicated by the model's weight (Figure 5C). Training the model with only 33 functional features resulted in unreliable accuracy (82%) and a high false-positive rate. However, the inclusion of “size” substantially improved the model's performance. Notably, the influence of “size” on viral pathogenicity lacks a discernible trend (Figure 6), resulting in predictive outcomes that tend towards a more stochastic distribution. Interaction analysis revealed that “viral counter signaling” has the strongest interaction with “size” (Figure 7). Even with these interactions, determining the direction of pathogenicity remains challenging. In summary, 'size' appears to fine-tune the model's final predictions. When combined with functional features, it facilitates a more accurate assessment of the likelihood of pathogenicity to humans.

In the validation results using an additional dataset, we identified four viruses with the potential to infect humans and induce diseases. Firstly, all sequence of Dabie bandavirus were predicted to be pathogenic. These viruses collected from ticks in Miyazaki Prefecture, Japan, exhibited a high degree of homology through phylogenetic analysis with a virus previously isolated from an SFTS patient, providing strong evidence for its potential pathogenicity (Sato et al., 2021). The model also predicted the pathogenicity of a strain of JEV, detected in mosquitoes in the Qinghai-Tibet region of China (Li et al., 2011). Despite the elevated altitude of the region, the presence of antibodies against the virus in both the indigenous population and swine suggests a localized occurrence of virus transmission, thereby challenging the initial presumption that the virus would not be prevalent at higher elevations. Tataguine virus (Kapuscinski et al., 2021), isolated from *Anopheles sp.* in Gambia, belongs to the *Peribunyaviridae*. While its pathogenicity to humans remains inconclusive, there is a high likelihood of infection symptoms if transmitted through hematophagous vectors. Among the 25 strains of Ebinur Lake Virus, one isolated from *Hyalomma marginatum* in the Volgograd region of Russia in 2023 was predicted to be capable of infecting humans. This virus, commonly found in China's prevalent vector host, *Culex modestus*, has been studied extensively for its ability to infect BALB/c mice, resulting in pronounced clinical symptoms (Zhao et al., 2020). Recent studies have substantiated the capacity of *Aedes aegypti* to serve as a vector for such viruses (C. Yang et al., 2022). Although antibodies have been detected in human serum samples, the lack of positive RT-PCR results prevents a conclusive determination of the virus's ability to infect humans and induce diseases (Xia et al., 2020). This aligns with the model's prediction, as the six viruses isolated from *Culex modestus*, included in the model, are unlikely to be pathogenic to humans.

This study endeavors to leverage machine learning methodologies for discerning overarching factors influencing the pathogenicity of hematophagous vector-borne viruses in humans. Our developed predictive model, focused on gene function, has successfully demonstrated the capability to predict virus pathogenicity in humans. However, it is crucial to acknowledge certain limitations in our study. In the global dataset of vector-borne viruses, there exists an uneven distribution, particularly with an overabundance of viruses such as DEV. This imbalance may result in an unavoidable bias that impacts the accuracy of the model.

Furthermore, the selected machine learning algorithms, while effective, may not match the efficacy of neural networks, posing challenges in optimizing for the current abundance of data. Notably, variations in blood-feeding habits among hematophagous vectors were not considered, which can significantly contribute to the spread of viruses. Different vector species may exhibit distinct preferences and behaviors in their blood-feeding patterns, influencing the transmission dynamics of viruses. Future research should incorporate these behavioral nuances to provide a more comprehensive understanding of virus dissemination. In summary, our model provides a novel perspective and serves as a valuable tool for the further analysis of virus sequences, providing effective information for the monitoring and early warning of hematophagous arthropod vector-borne transmission.

In this investigation, our primary objective is to discern both macroscopic and microscopic factors influencing the risk of human pathogenicity in hematophagous vector-borne viruses. Employing ensemble learning standpoint, we uncovered key characteristics associated with viral pathogenicity from an epidemic perspective. Simultaneously, we delved into pivotal functional features impacting human pathogenicity at a molecular level, with a specific focus on the functional aspects of viral sequences. Moreover, we deploy our developed model to forecast the human infectivity of viral sequences within an additional validation dataset. The model's performance in predicting the pathogenicity of these viruses at the genetic level not only enriches our comprehension of established and emerging virus risks but also broadens the scope of hematophagous arthropods detection. Importantly, it contributes substantively to the mitigation of present and future risks associated with vector-borne diseases.

Acknowledgement

The authors acknowledge the global open dataset shared by Huang et al and Xuan Li for assistance with additional data collection. The laboratory is funded by a grant from National Key Research and Development Program of China (2019YFC1200501). The funders had no role in study design, data collection and interpretation, or the decision to submit the work for publication.

Data availability

The data supporting the findings of this study are available upon reasonable request from the author. Researchers interested in accessing the dataset for further exploration or verification are encouraged to contact Huakai Hu at hhyu98@163.com for assistance. We are committed to promoting transparency and collaboration in scientific research, and we welcome inquiries regarding the data underlying our published results.

Author contribution

Huakai Hu, Idea Generation, Data Curation and Transformation, Model Development and validation, Visualization, Writing - original; Chaoying Zhao, Conceptualization, Methodology, Writing - original draft and review; Jiali Chen, Conceptualization, Methodology, Writing – review and editing; Meiling Jin, Conceptualization, Writing – review and editing; Hua Shi, Conceptualization, Writing – review and editing; Jinpeng Guo, Project administration, Writing – review and editing; Changjun Wang, Conceptualization, Methodology, Writing – review and editing; Yong Chen, Supervision, Funding acquisition, Project administration, Writing – review and editing;

References

Balaji, A., Kille, B., Kappell, A. D., Godbold, G. D., Diep, M., Elworth, R. A. L., Qian, Z., Albin, D., Nasko, D. J., Shah, N., Pop, M., Segarra, S., Ternus, K. L., & Treangen, T. J. (2022). SeqScreen: Accurate and sensitive functional screening of pathogenic sequences via ensemble learning. *Genome Biology*, *23*(1), 133. https://doi.org/10.1186/s13059-022-02695-x

Bartoszewicz, J. M., Genske, U., & Renard, B. Y. (2021). Deep learning-based real-time detection of novel pathogens during sequencing. *Briefings in Bioinformatics*, *22*(6), bbab269. https://doi.org/10.1093/bib/bbab269

Bartoszewicz, J. M., Seidel, A., & Renard, B. Y. (2021). Interpretable detection of novel human viruses from genome sequencing data. *NAR Genomics and Bioinformatics*, *3*(1), lqab004. https://doi.org/10.1093/nargab/lqab004

Batson, J., Dudas, G., Haas-Stapleton, E., Kistler, A. L., Li, L. M., Logan, P., Ratnasiri, K., & Retallack, H. (2021). Single mosquito metatranscriptomics identifies vectors, emerging pathogens and reservoirs in one assay. *eLife*, *10*, e68353. https://doi.org/10.7554/eLife.68353

Begum, F., Das, S., Mukherjee, D., & Ray, U. (2019). Hijacking the Host Immune Cells by Dengue Virus: Molecular Interplay of Receptors and Dengue Virus Envelope. *Microorganisms*, *7*(9), Article 9. https://doi.org/10.3390/microorganisms7090323

Behl, A., Nair, A., Mohagaonkar, S., Yadav, P., Gambhir, K., Tyagi, N., Sharma, R. K., Butola, B. S., & Sharma, N. (2022). Threat, challenges, and preparedness for future pandemics: A descriptive review of phylogenetic analysis based predictions. *Infection, Genetics and Evolution: Journal of Molecular Epidemiology and Evolutionary Genetics in Infectious Diseases*, *98*, 105217. https://doi.org/10.1016/j.meegid.2022.105217

Birnberg, L., Temmam, S., Aranda, C., Correa-Fiz, F., Talavera, S., Bigot, T., Eloit, M., & Busquets, N. (2020). Viromics on Honey-Baited FTA Cards as a New Tool f

or the Detection of Circulating Viruses in Mosquitoes. *Viruses*, *12*(3), 274. https://doi.org/10.3390/v12030274

Brinkmann, A., Nitsche, A., & Kohl, C. (2016). Viral Metagenomics on Blood-Feeding Arthropods as a Tool for Human Disease Surveillance. *International Journal of Molecular Sciences*, *17*(10), Article 10. https://doi.org/10.3390/ijms17101743

Calisher, C. H., & Higgs, S. (2018). The Discovery of Arthropod-Specific Viruses in Hematophagous Arthropods: An Open Door to Understanding the Mechanisms of Arbovirus and Arthropod Evolution? *Annual Review of Entomology*, *63*(1), 87–103. https://doi.org/10.1146/annurev-ento-020117-043033

Chala, B., & Hamde, F. (2021). Emerging and Re-emerging Vector-Borne Infectious Diseases and the Challenges for Control: A Review. *Frontiers in Public Health*, *9*. https://www.frontiersin.org/articles/10.3389/fpubh.2021.715759

Ciota, A. T., & Keyel, A. C. (2019). The Role of Temperature in Transmission of Zoonotic Arboviruses. *Viruses*, *11*(11), Article 11. https://doi.org/10.3390/v11111013

Coatsworth, H., Bozic, J., Carrillo, J., Buckner, E. A., Rivers, A. R., Dinglasan, R. R., & Mathias, D. K. (2022). Intrinsic variation in the vertically transmitted core virome of the mosquito Aedes aegypti. *Molecular Ecology*, *31*(9), 2545–2561. https://doi.org/10.1111/mec.16412

Conway, M. J., Colpitts, T. M., & Fikrig, E. (2014). Role of the Vector in Arbovirus Transmission. *Annual Review of Virology*, *1*(1), 71–88. https://doi.org/10.1146/annurev-virology-031413-085513

Costa, V. V., Fagundes, C. T., Souza, D. G., & Teixeira, M. M. (2013). Inflammatory and Innate Immune Responses in Dengue Infection: Protection versus Disease Induction. *The American Journal of Pathology*, *182*(6), 1950–1961. https://doi.org/10.1016/j.ajpath.2013.02.027

Cruz-Oliveira, C., Freire, J. M., Conceição, T. M., Higa, L. M., Castanho, M. A. R.

B., & Da Poian, A. T. (2015). Receptors and routes of dengue virus entry into the host cells. *FEMS Microbiology Reviews*, *39*(2), 155–170. https://doi.org/10.1093/femsre/fuu004

Cuthbert, R. N., Darriet, F., Chabrerie, O., Lenoir, J., Courchamp, F., Claeys, C., Robert, V., Jourdain, F., Ulmer, R., Diagne, C., Ayala, D., Simard, F., Morand, S., & Renault, D. (2023). Invasive hematophagous arthropods and associated diseases in a changing world. *Parasites & Vectors*, *16*(1), 291. https://doi.org/10.1186/s13071-023-05887-x

Fang, Z., Tan, J., Wu, S., Li, M., Xu, C., Xie, Z., & Zhu, H. (2019). PPR-Meta: A tool for identifying phages and plasmids from metagenomic fragments using deep learning. *GigaScience*, *8*(6), giz066. https://doi.org/10.1093/gigascience/giz066

Faustino, A. F., Martins, A. S., Karguth, N., Artilheiro, V., Enguita, F. J., Ricardo, J. C., Santos, N. C., & Martins, I. C. (2019). Structural and Functional Properties of the Capsid Protein of Dengue and Related Flavivirus. *International Journal of Molecular Sciences*, *20*(16), Article 16. https://doi.org/10.3390/ijms20163870

Forrester, N. L., Coffey, L. L., & Weaver, S. C. (2014). Arboviral Bottlenecks and Challenges to Maintaining Diversity and Fitness during Mosquito Transmission. *Viruses*, *6*(10), Article 10. https://doi.org/10.3390/v6103991

Fournet, N., Voiry, N., Rozenberg, J., Bassi, C., Cassonnet, C., Karch, A., Durand, G., Grard, G., Modenesi, G., Lakoussan, S.-B., Tayliam, N., Zatta, M., Gallien, S., investigation team, Noël, H., Brichler, S., & Tarantola, A. (2023). A cluster of autochthonous dengue transmission in the Paris region—Detection, epidemiology and control measures, France, October 2023. *Euro Surveillance: Bulletin Europeen Sur Les Maladies Transmissibles = European Communicable Disease Bulletin*, *28*(49). https://doi.org/10.2807/1560-7917.ES.2023.28.49.2300641

García-Romero, C., Carrillo Bilbao, G. A., Navarro, J.-C., Martin-Solano, S., & Saegerman, C. (2023). Arboviruses in Mammals in the Neotropics: A Systematic Review t

o Strengthen Epidemiological Monitoring Strategies and Conservation Medicine. *Viruses*, *15*(2), Article 2. https://doi.org/10.3390/v15020417

Geoghegan, J. L., & Holmes, E. C. (2018). The phylogenomics of evolving virus virulence. *Nature Reviews Genetics*, *19*(12), Article 12. https://doi.org/10.1038/s41576-018-0055-5

Golnar, A. J., Turell, M. J., LaBeaud, A. D., Kading, R. C., & Hamer, G. L. (2014). Predicting the Mosquito Species and Vertebrate Species Involved in the Theoretical Transmission of Rift Valley Fever Virus in the United States. *PLOS Neglected Tropical Diseases*, *8*(9), e3163. https://doi.org/10.1371/journal.pntd.0003163

Gould, E., Pettersson, J., Higgs, S., Charrel, R., & de Lamballerie, X. (2017). Emerging arboviruses: Why today? *One Health*, *4*, 1–13. https://doi.org/10.1016/j.onehlt.2017.06.001

Hasan, S. S., Miller, A., Sapparapu, G., Fernandez, E., Klose, T., Long, F., Fokine, A., Porta, J. C., Jiang, W., Diamond, M. S., Crowe Jr., J. E., Kuhn, R. J., & Rossmann, M. G. (2017). A human antibody against Zika virus crosslinks the E protein to prevent infection. *Nature Communications*, *8*(1), Article 1. https://doi.org/10.1038/ncomms14722

Hermanns, K., Marklewitz, M., Zirkel, F., Kopp, A., Kramer-Schadt, S., & Junglen, S. (2023). Mosquito community composition shapes virus prevalence patterns along anthropogenic disturbance gradients. *eLife*, *12*, e66550. https://doi.org/10.7554/eLife.66550

Huang, Y., Wang, S., Liu, H., Atoni, E., Wang, F., Chen, W., Li, Z., Rodriguez, S., Yuan, Z., Ming, Z., & Xia, H. (2023). A global dataset of sequence, diversity and biosafety recommendation of arbovirus and arthropod-specific virus. *Scientific Data*, *10*(1), 305. https://doi.org/10.1038/s41597-023-02226-8

Huang, Y.-J. S., Higgs, S., & Vanlandingham, D. L. (2019a). Arbovirus-Mosquito Vect

or-Host Interactions and the Impact on Transmission and Disease Pathogenesis of Arboviruses. *Frontiers in Microbiology*, *10*. https://www.frontiersin.org/articles/10.3389/fmicb.2019.00022

Huang, Y.-J. S., Higgs, S., & Vanlandingham, D. L. (2019b). Emergence and re-emergence of mosquito-borne arboviruses. *Current Opinion in Virology*, *34*, 104–109. https://doi.org/10.1016/j.coviro.2019.01.001

Kampen, H., & Werner, D. (2014). Out of the bush: The Asian bush mosquito Aedes japonicus japonicus (Theobald, 1901) (Diptera, Culicidae) becomes invasive. *Parasites & Vectors*, *7*, 59. https://doi.org/10.1186/1756-3305-7-59

Kapuscinski, M. L., Bergren, N. A., Russell, B. J., Lee, J. S., Borland, E. M., Hartman, D. A., King, D. C., Hughes, H. R., Burkhalter, K. L., Kading, R. C., & Stenglein, M. D. (2021). Genomic characterization of 99 viruses from the bunyavirus families Nairoviridae, Peribunyaviridae, and Phenuiviridae, including 35 previously unsequenced viruses. *PLoS Pathogens*, *17*(3), e1009315. https://doi.org/10.1371/journal.ppat.1009315

Khongwichit, S., Chuchaona, W., Vongpunsawad, S., & Poovorawan, Y. (2023). Molecular epidemiology, clinical analysis, and genetic characterization of Zika virus infections in Thailand (2020-2023). *Scientific Reports*, *13*(1), 21030. https://doi.org/10.1038/s41598-023-48508-4

King, C. A., Wegman, A. D., & Endy, T. P. (2020). Mobilization and Activation of the Innate Immune Response to Dengue Virus. *Frontiers in Cellular and Infection Microbiology*, *10*. https://www.frontiersin.org/articles/10.3389/fcimb.2020.574417

Lewis, W. H., Tahon, G., Geesink, P., Sousa, D. Z., & Ettema, T. J. G. (2021). Innovations to culturing the uncultured microbial majority. *Nature Reviews. Microbiology*, *19*(4), 225–240. https://doi.org/10.1038/s41579-020-00458-8

Li, Y.-X., Li, M.-H., Fu, S.-H., Chen, W.-X., Liu, Q.-Y., Zhang, H.-L., Da, W., Hu, S.

-L., La Mu, S. D., Bai, J., Yin, Z.-D., Jiang, H.-Y., Guo, Y.-H., Ji, D. Z. D., Xu, H.-M., Li, G., Mu, G. G. C., Luo, H.-M., Wang, J.-L., … Liang, G.-D. (2011). Japanese Encephalitis, Tibet, China. *Emerging Infectious Diseases*, *17*(5), 934–936. https://doi.org/10.3201/eid1705.101417

Martins, A. S., Carvalho, F. A., Faustino, A. F., Martins, I. C., & Santos, N. C. (2019). West Nile Virus Capsid Protein Interacts With Biologically Relevant Host Lipid Systems. *Frontiers in Cellular and Infection Microbiology*, *9*. https://www.frontiersin.org/articles/10.3389/fcimb.2019.00008

Mollentze, N., Babayan, S. A., & Streicker, D. G. (2021). Identifying and prioritizing potential human-infecting viruses from their genome sequences. *PLOS Biology*, *19*(9), e3001390. https://doi.org/10.1371/journal.pbio.3001390

Ni, X.-B., Cui, X.-M., Liu, J.-Y., Ye, R.-Z., Wu, Y.-Q., Jiang, J.-F., Sun, Y., Wang, Q., Shum, M. H.-H., Chang, Q.-C., Zhao, L., Han, X.-H., Ma, K., Shen, S.-J., Zhang, M.-Z., Guo, W.-B., Zhu, J.-G., Zhan, L., Li, L.-J., … Cao, W.-C. (2023). Metavirome of 31 tick species provides a compendium of 1,801 RNA virus genomes. *Nature Microbiology*, *8*(1), 162–173. https://doi.org/10.1038/s41564-022-01275-w

Nouri, S., Matsumura, E. E., Kuo, Y.-W., & Falk, B. W. (2018). Insect-specific viruses: From discovery to potential translational applications. *Current Opinion in Virology*, *33*, 33–41. https://doi.org/10.1016/j.coviro.2018.07.006

Roth, G. A., Abate, D., Abate, K. H., Abay, S. M., Abbafati, C., Abbasi, N., Abbastabar, H., Abd-Allah, F., Abdela, J., Abdelalim, A., Abdollahpour, I., Abdulkader, R. S., Abebe, H. T., Abebe, M., Abebe, Z., Abejie, A. N., Abera, S. F., Abil, O. Z., Abraha, H. N., … Murray, C. J. L. (2018). Global, regional, and national age-sex-specific mortality for 282 causes of death in 195 countries and territories, 1980–2017: A systematic analysis for the Global Burden of Disease Study 2017. *The Lancet*, *392*(10159), 1736–1788. https://doi.org/10.1016/S0140-6736(18)32203-7

Sato, Y., Mekata, H., Sudaryatma, P. E., Kirino, Y., Yamamoto, S., Ando, S., Sugimot

o, T., & Okabayashi, T. (2021). Isolation of Severe Fever with Thrombocytopenia Syndrome Virus from Various Tick Species in Area with Human Severe Fever with Thrombocytopenia Syndrome Cases. *Vector-Borne and Zoonotic Diseases*, *21*(5), 378–384. https://doi.org/10.1089/vbz.2020.2720

Sen, R., Nayak, L., & De, R. K. (2016). A review on host–pathogen interactions: Classification and prediction. *European Journal of Clinical Microbiology & Infectious Diseases*, *35*(10), 1581–1599. https://doi.org/10.1007/s10096-016-2716-7

Stephenson, E. B., Murphy, A. K., Jansen, C. C., Peel, A. J., & McCallum, H. (2019). Interpreting mosquito feeding patterns in Australia through an ecological lens: An analysis of blood meal studies. *Parasites & Vectors*, *12*(1), 156. https://doi.org/10.1186/s13071-019-3405-z

Tabachnick, W. J. (2016). Climate Change and the Arboviruses: Lessons from the Evolution of the Dengue and Yellow Fever Viruses. *Annual Review of Virology*, *3*(1), 125–145. https://doi.org/10.1146/annurev-virology-110615-035630

Touray, M., Bakirci, S., Ulug, D., Gulsen, S. H., Cimen, H., Yavasoglu, S. I., Simsek, F. M., Ertabaklar, H., Ozbel, Y., & Hazir, S. (2023). Arthropod vectors of disease agents: Their role in public and veterinary health in Turkiye and their control measures. *Acta Tropica*, *243*, 106893. https://doi.org/10.1016/j.actatropica.2023.106893

Viglietta, M., Bellone, R., Blisnick, A. A., & Failloux, A.-B. (2021). Vector Specificity of Arbovirus Transmission. *Frontiers in Microbiology*, *12*. https://www.frontiersin.org/articles/10.3389/fmicb.2021.773211

Vogels, C. B. F., Rückert, C., Cavany, S. M., Perkins, T. A., Ebel, G. D., & Grubaugh, N. D. (2019). Arbovirus coinfection and co-transmission: A neglected public health concern? *PLOS Biology*, *17*(1), e3000130. https://doi.org/10.1371/journal.pbio.3000130

Weaver, S. C., Charlier, C., Vasilakis, N., & Lecuit, M. (2018). Zika, Chikungunya, an

d Other Emerging Vector-Borne Viral Diseases. *Annual Review of Medicine*, *69*, 395–408. https://doi.org/10.1146/annurev-med-050715-105122

Weissenböck, H., Hubálek, Z., Bakonyi, T., & Nowotny, N. (2010). Zoonotic mosquito-borne flaviviruses: Worldwide presence of agents with proven pathogenicity and potential candidates of future emerging diseases. *Veterinary Microbiology*, *140*(3–4), 271–280. https://doi.org/10.1016/j.vetmic.2009.08.025

Wu, Z., Zhang, M., Zhang, Y., Lu, K., Zhu, W., Feng, S., Qi, J., & Niu, G. (2023). Jingmen tick virus: An emerging arbovirus with a global threat. *mSphere*, *8*(5), e00281-23. https://doi.org/10.1128/msphere.00281-23

Xia, H., Liu, R., Zhao, L., Sun, X., Zheng, Z., Atoni, E., Hu, X., Zhang, B., Zhang, G., & Yuan, Z. (2020). Characterization of Ebinur Lake Virus and Its Human Seroprevalence at the China–Kazakhstan Border. *Frontiers in Microbiology*, *10*. https://www.frontiersin.org/articles/10.3389/fmicb.2019.03111

Yang, C., Wang, F., Huang, D., Ma, H., Zhao, L., Zhang, G., Li, H., Han, Q., Bente, D., Salazar, F. V., Yuan, Z., & Xia, H. (2022). Vector competence and immune response of Aedes aegypti for Ebinur Lake virus, a newly classified mosquito-borne orthobunyavirus. *PLoS Neglected Tropical Diseases*, *16*(7), e0010642. https://doi.org/10.1371/journal.pntd.0010642

Yang, X., Qin, S., Liu, X., Zhang, N., Chen, J., Jin, M., Liu, F., Wang, Y., Guo, J., Shi, H., Wang, C., & Chen, Y. (2023). Meta-Viromic Sequencing Reveals Virome Characteristics of Mosquitoes and *Culicoides* on Zhoushan Island, China. *Microbiology Spectrum*, e02688-22. https://doi.org/10.1128/spectrum.02688-22

Zaid, A., Burt, F. J., Liu, X., Poo, Y. S., Zandi, K., Suhrbier, A., Weaver, S. C., Texeira, M. M., & Mahalingam, S. (2021). Arthritogenic alphaviruses: Epidemiological and clinical perspective on emerging arboviruses. *The Lancet Infectious Diseases*, *21*(5), e123–e133. https://doi.org/10.1016/S1473-3099(20)30491-6

Zhang, Z., Cai, Z., Tan, Z., Lu, C., Jiang, T., Zhang, G., & Peng, Y. (2019). Rapid identification of human-infecting viruses. *Transboundary and Emerging Diseases*, *66* (6), 2517–2522. https://doi.org/10.1111/tbed.13314

Zhao, L., Luo, H., Huang, D., Yu, P., Dong, Q., Mwaliko, C., Atoni, E., Nyaruaba, R., Yuan, J., Zhang, G., Bente, D., Yuan, Z., & Xia, H. (2020). Pathogenesis and Immune Response of Ebinur Lake Virus: A Newly Identified Orthobunyavirus That Exhibited Strong Virulence in Mice. *Frontiers in Microbiology*, *11*, 625661. https://doi.org/10.3389/fmicb.2020.625661

Zhao, L., Yu, P., Shi, C., Jia, L., Evans, A., Wang, X., Wu, Q., Xiong, G., Ming, Z., Salazar, F., Agwanda, B., Bente, D., Wang, F., Liu, D., Yuan, Z., & Xia, H. (2022). *Global mosquito virome profiling and mosquito spatial diffusion pathways revealed by marker-viruses* [Preprint]. Microbiology. https://doi.org/10.1101/2022.09.24.509300

Table 1: Summary of Epidemiological characteristics in regression model. A detailed summary of the 37 epidemiological characteristics considered in our regression model.

Classification	Name of characteristics	Detailed description of characteristics
virus	Virus_Group	Arboviruses or ISVs
	virus	virus name
	vi_G	virus genus
	vi_F	virus family
	Count	virus counts
	baltimore	virus baltimore classification
vector hosts	vector_G	vector genus
	vector_F	vector family
	vector_O	vector order
	vector_C	vector class
	continent	vector continent
	country	vector country
	climate	vector climate
	cross_vector_G	Counts of cross-vector host genera
	cross_vector_F	Counts of cross-vector host families
	cross_vector_O	Counts of cross-vector host orders
	cross_vector_C	Counts of cross-vector host classes
	total_vector	Total counts of cross-vector hosts
	vector_G_T	Total counts of cross-vector host genera
	vector_F_T	Total counts of cross-vector host families
	vector_O_T	Total counts of cross-vector host orders
	vector_C_T	Total counts of cross-vector host classes
non-vector hosts	Aves	Aves host
	Carnivora	Carnivora host
	Rodentia	Rodentia host
	Chiroptera	Chiroptera host
	Primates	Primates host
	homo	homo host
	Didelphimorphia	Didelphimorphia host
	Artiodactyla	Artiodactyla host
	Perissodactyla	Perissodactyla host
	Eulipotyphla	Eulipotyphla host
	Reptilia	Reptilia host
	Lagomorpha	Lagomorpha host
	Anura	Anura host
	Pilosa	Pilosa host
	Diprotodontia	Diprotodontia host
	cross_host	total cross non-vectors host
	homo_infected	Whether or not homuns are infectious

Table 2: Summary of FunSoCs annotation results from SeqScreen. Counts and definitions of 10 distinct FunSoCs identified in this dataset.

FunSoC title	Counts	FunSoC definition
Viral adhesion	44482	Mediates viral adherence to host cells
Viral counter signaling	35256	Viral suppression of host immune signaling within host cells to avoid inflammatory responses
Host xenophagy	33656	Target host xenophagy/autophagy
Viral invasion	4376	Mediates viral invasion into host cell
Host transcription	949	Target host transcription to inhibit or activate
Host ubiquitin	880	Target host ubiquitination machinery
Host cell death	802	Target host apoptotic cell death pathways either to inhibit or activate
Resist complement	144	Enable resistance from host complement components
Antibiotic resistance	4	Counters the effect of antibiotics administered to inhibit the growth or vital functioning of bacterial or eukaryotic parasites.
Induce inflammation	1	Directly activate host inflammatory pathways to cause damage

Figure 1: Hematophagous arthropod vector and non-vector hosts characteristics in the dataset. (A) The global distribution and quantity of blood-sucking vectors and their carriers. (B) The number of vector hosts, the continents where they are located, and the types of hostile weather conditions where they are found. (C) The characteristics of the number of non-vector hosts. (D) Viruses transmitted across non-vector hosts Quantity. The abscissa is the number across non-vector hosts, and the ordinate is the total number of viruses.

Figure 2: Relative importance of different macroscopic characteristics in the regression model. The weight contributions of diverse epidemiological features of viruses in the regression model to human pathogenicity.

Figure 3: Hosts distribution in viral functions annotation. The distribution of hosts for known non-pathogenic viruses to humans (A) and known human-pathogenic viruses (B). The actual counts of viruses are converted to percentage representations in their respective sections of the chart.

Figure 4: Metrics of comprehensive assessment of model performance. The utilization of ROC Curve (A)and Confusion Matrix (B) for assessing the performance of the model.

Figure 5: Ranking of metrics presented in the XGBoost functional annotations model. Within the results of the XGBoost model, the functional feature importance outcomes of gain (A), cover (B), and weight (C) are separately obtained. These three metrics collectively reflect the relative significance in determining the pathogenicity of hematophagous arthropod vector-borne viruses.

Figure 6: The collective impact of viral function annotations on pathogeniciy prediction analyzed through SHAP. Providing a comprehensive overview of how various viral function annotations collectively contribute to the model's predictions regarding pathogenicity.

Figure 7: Detailed analysis of the interactions among crucial features in pathogenicity prediction models through SHAP. The interactions examined include those between viral sequence size and viral adhesion (A), host xenophagy and viral adhesion (B), as well as viral invasion and viral counter signaling (C). These analyses contribute to a deeper understanding of the combined influence of these features on pathogenicity predictions.

Supplementary Information

Table supplement 1: Family for hematophagous arthropod vector

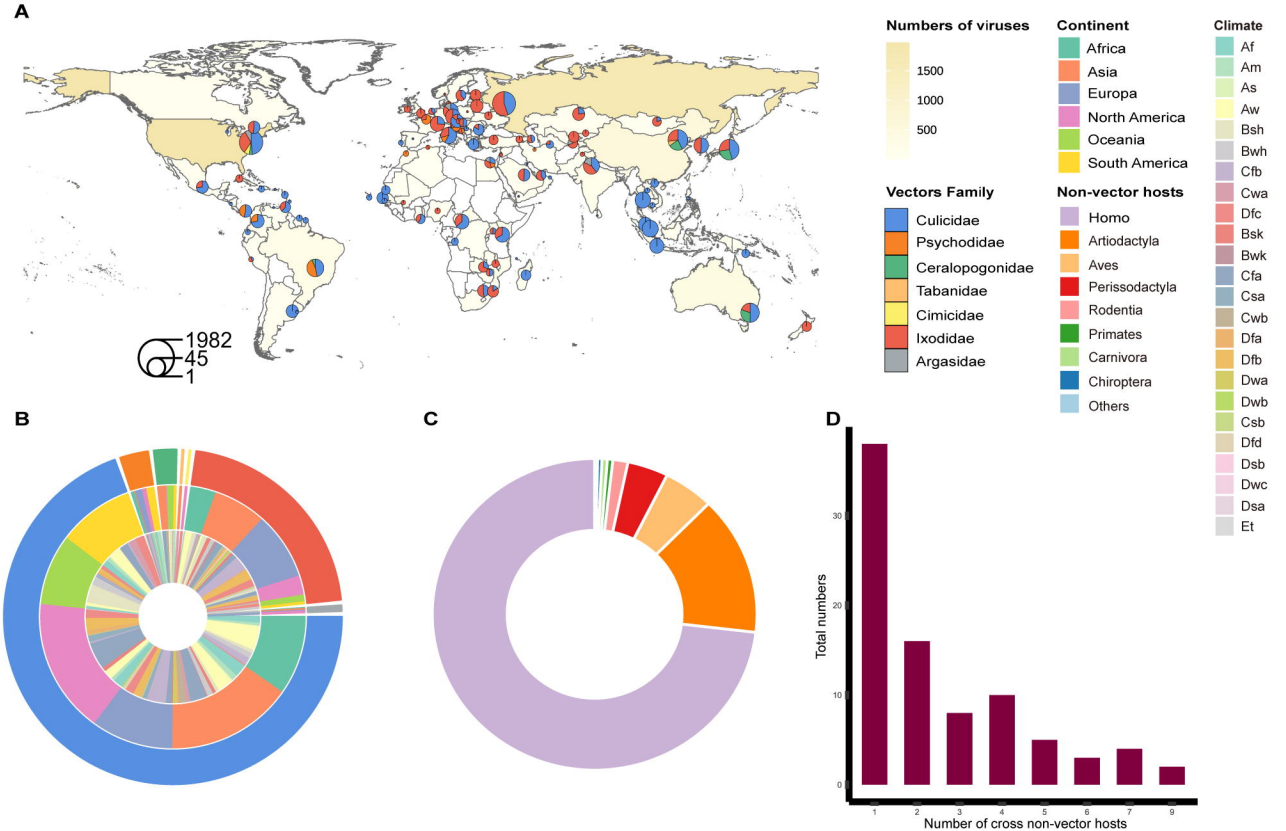
Family
Culicidae
Phlebotominae
Ceratopogonidae
Simuliidae
Tabanidae
Cimicidae
Ixodidae
Argasidae
Stenoponiidae
Phthiraptera

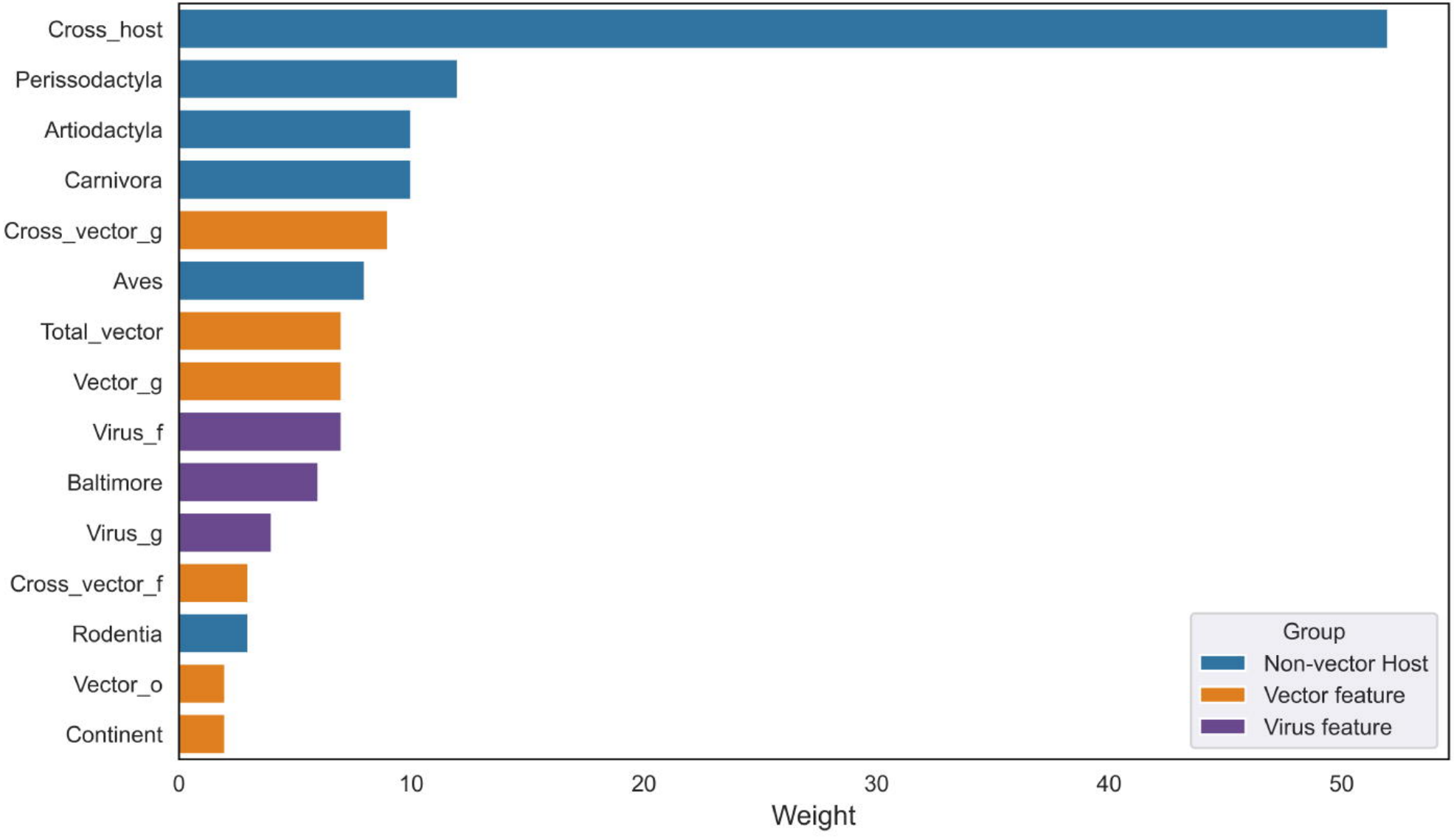
Table supplement 2: Hyperparameter settings for the XGBoost regression model. Optimized parameter settings for the XGBoost regression model obtained through rigorous experimentation and fine-tuning.

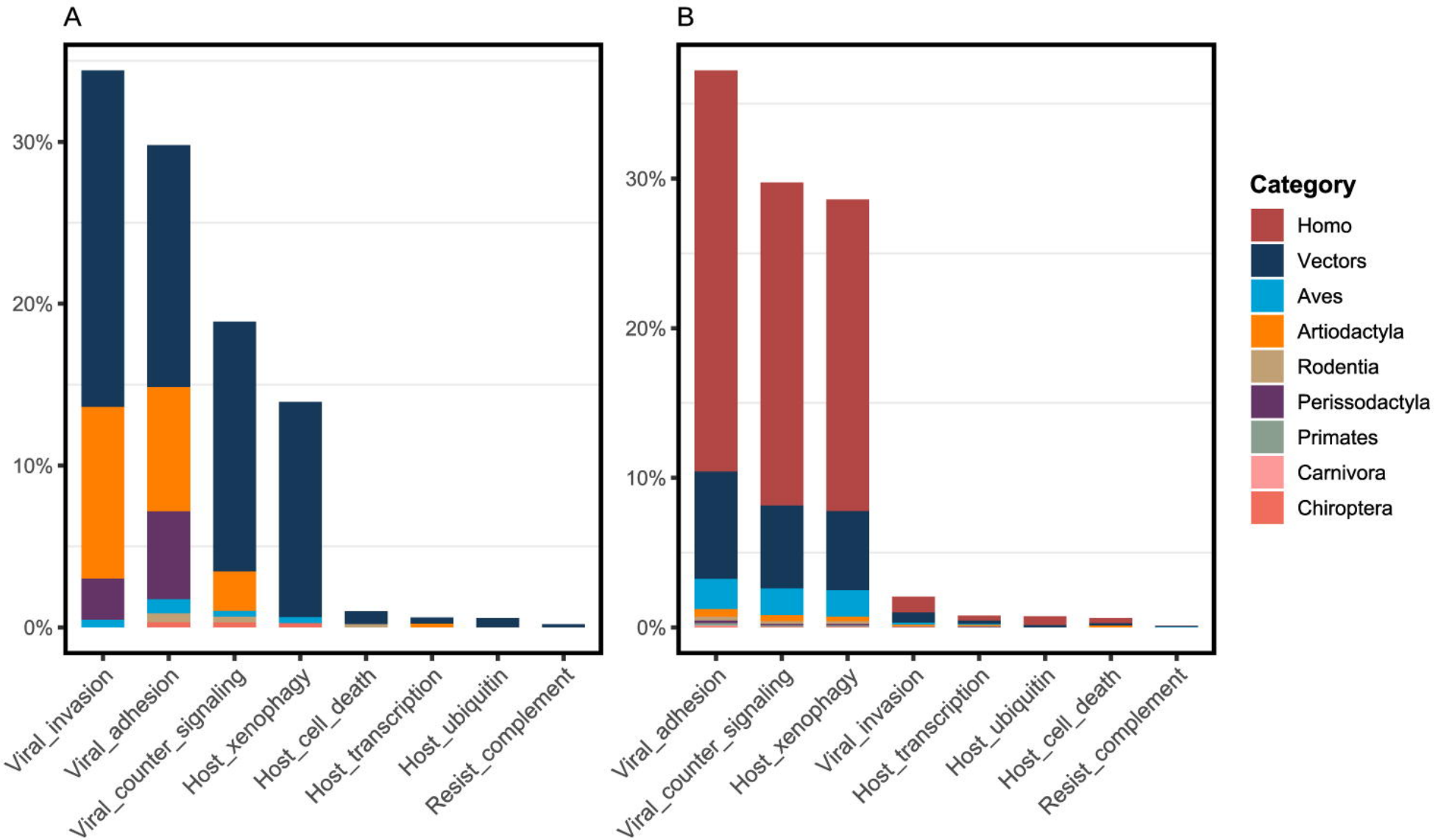
Hyperparameter	Value
booster	dart
eta	0.15
max_depth	3
subsample	0.7
objective	reg:logistic
tree_method	exact
max_cat_threshold	20
eval_metric	["logloss", "auc", 'error']

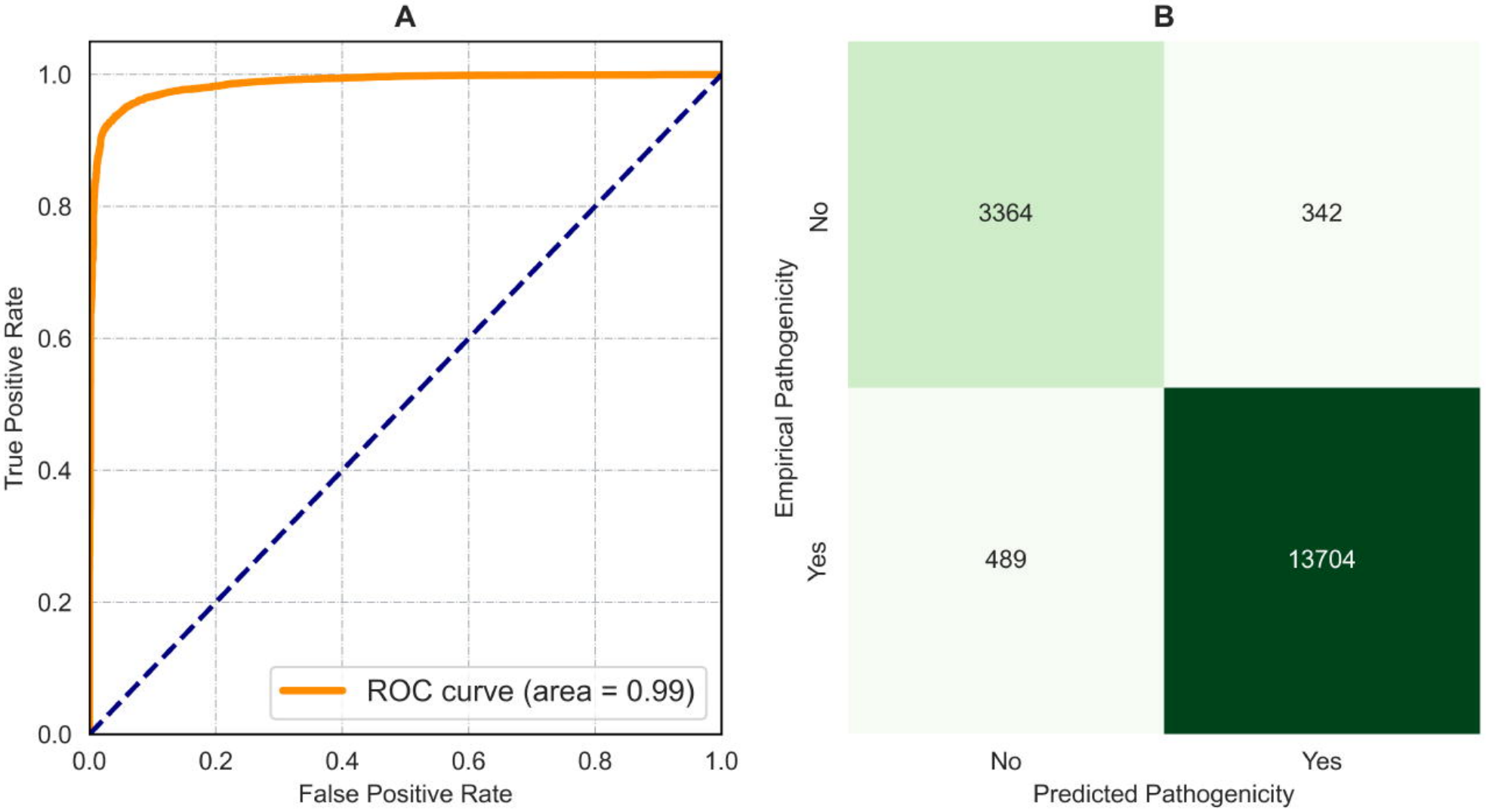
Table supplement 3: Hyperparameter settings for the XGBoost classification model. Optimized parameter settings for the XGBoost classification model obtained through rigorous experimentation and fine-tuning.

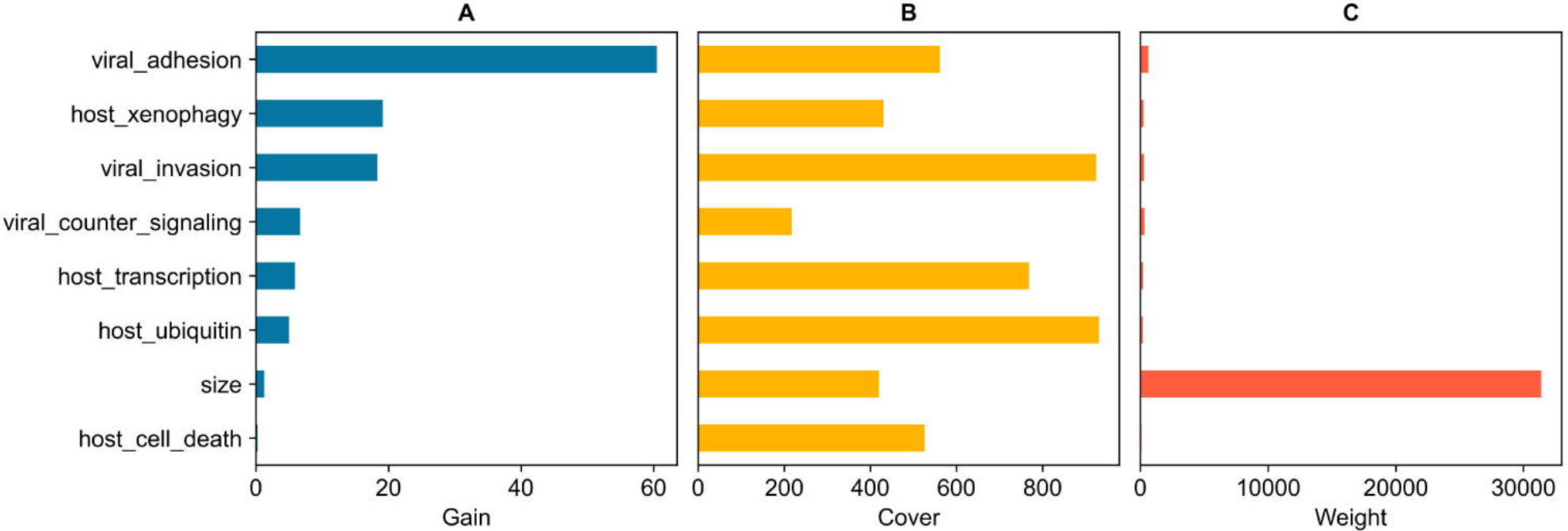
Hyperparameter	Value
objective	binary:logistic
tree_method	eaxct
scale_pos_weight	0.26
eta	0.15

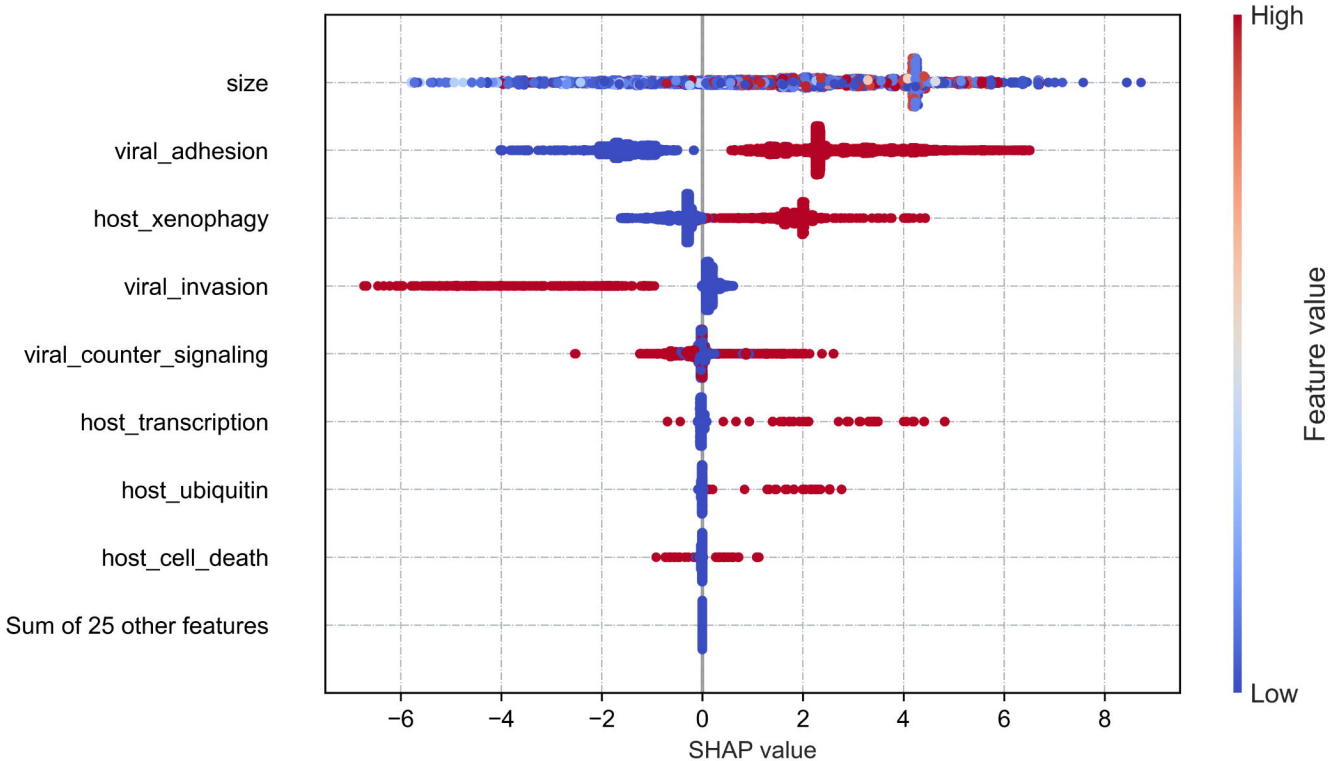

A
1982
45
1
Numbers of viruses
1500
1000
500
Continent
Africa
Asia
Europa
North America
Oceania
South America
Climate
Af
Am
As
Aw
Bsh
Bwh
Cfb
Cwa
Dfc
Bsk
Bwk
Cfa
Csa
Cwb
Dfa
Dfb
Dwa
Dwb
Csb
Dfd
Dsb
Dwc
Dsa
Et
Vectors Family
Culicidae
Psychodidae
Ceralopogonidae
Tabanidae
Cimicidae
Ixodidae
Argasidae
Non-vector hosts
Homo
Artiodactyla
Aves
Perissodactyla
Rodentia
Primates
Carnivora
Chiroptera
Others
B
C
D
Total numbers
0
10
20
30
1
2
3
4
5
6
7
9
Number of cross non-vector hosts

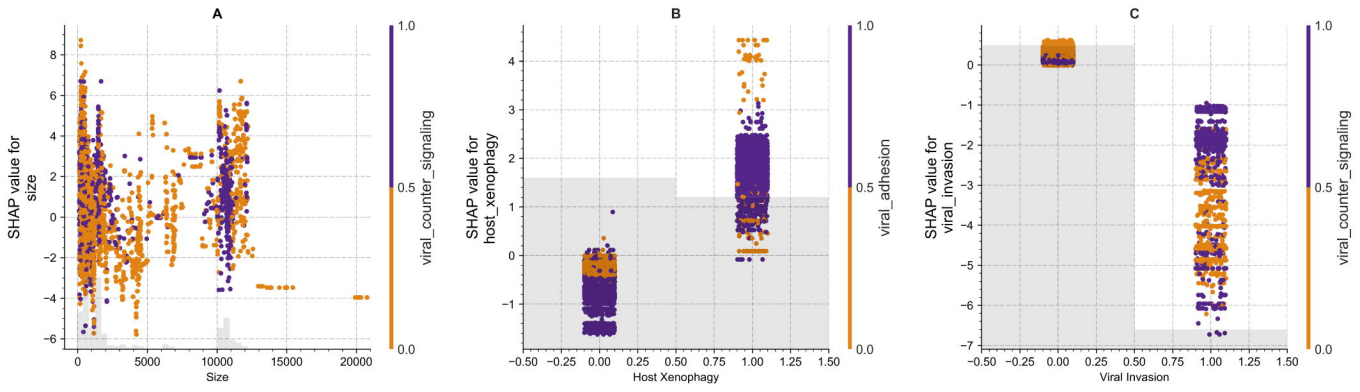

Cross_host
Perissodactyla
Artiodactyla
Carnivora
Cross_vector_g
Aves
Total_vector
Vector_g
Virus_f
Baltimore
Virus_g
Cross_vector_f
Rodentia
Vector_o
Continent
0
10
20
30
40
50
Weight
Group
Non-vector Host
Vector feature
Virus feature


A
B
30%
20%
10%
0%
Viral_invasion
Viral_adhesion
Viral_counter_signaling
Host_xenophagy
Host_cell_death
Host_transcription
Host_ubiquitin
Resist_complement
Category
Homo
Vectors
Aves
Artiodactyla
Rodentia
Perissodactyla
Primates
Carnivora
Chiroptera


A
1.0
0.8
0.6
0.4
0.2
0.0
True Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
False Positive Rate
ROC curve (area = 0.99)
B
Empirical Pathogenicity
No
Yes
3364
342
489
13704
No
Yes
Predicted Pathogenicity


A
B
C
viral_adhesion
host_xenophagy
viral_invasion
viral_counter_signaling
host_transcription
host_ubiquitin
size
host_cell_death
0
20
40
60
Gain
0
200
400
600
800
Cover
0
10000
20000
30000
Weight


size
viral_adhesion
host_xenophagy
viral_invasion
viral_counter_signaling
host_transcription
host_ubiquitin
host_cell_death
Sum of 25 other features
−6
−4
−2
0
2
4
6
8
SHAP value
High
Feature value
Low


A
SHAP value for size
Size
viral_counter_signaling
B
SHAP value for host_xenophagy
Host Xenophagy
viral_adhesion
C
SHAP value for viral_invasion
Viral Invasion
viral_counter_signaling