Viral genomic features predict orthopoxvirus reservoir hosts

Katie K. Tseng[a], Heather Koehler[b], Daniel J. Becker[c], Rory Gibb[d,e], Colin J. Carlson[f], Maria del Pilar Fernandez[a*], and Stephanie N. Seifert[a*]

[a] Paul G. Allen School for Global Health, Washington State University, Pullman, WA, USA
[b] School of Molecular Biosciences, Washington State University, Pullman, WA, USA
[c] Department of Biology, School of Biological Sciences, University of Oklahoma, Norman, OK, USA
[d] Centre for Biodiversity and Environment Research, Department of Genetics, Evolution and Environment, University College London, London, UK
[e] People & Nature Lab, UCL East, University College London, Stratford, London, UK
[f] Center for Global Health Science and Security, Georgetown University, Washington, DC, USA

***Corresponding Authors:** Stephanie N. Seifert and M. Pilar Fernandez.
Email: stephanie.seifert@wsu.edu and pilar.fernandez@wsu.edu
Author Contributions: K.K.T, P.F. and S.N.S. designed research; K.K.T., H.K., M.P.F. and S.N.S. performed research; D.J.B., R.G. and C.J.C. contributed to analytic tools; K.K.T, H.K., M.P.F. and S.N.S. analyzed data; and K.K.T., H.K., M.P.F. and S.N.S. wrote the paper. All authors reviewed the paper.
Competing Interest Statement: The authors declare no competing interest.
Classification: Biological Sciences and Ecology.
Keywords: Infectious diseases | prediction | machine learning | boosted regression trees | orthopoxviruses, Mpox

This PDF file includes:
Main Text
Figures 1 to 3
Tables 1 to 1

Abstract

Orthopoxviruses (OPVs), including the causative agents of smallpox and mpox have led to devastating outbreaks in human populations worldwide. However, the discontinuation of smallpox vaccination, which also provides cross-protection against related OPVs, has diminished global immunity to OPVs more broadly. We apply machine learning models incorporating both host ecological and viral genomic features to predict likely reservoirs of OPVs. We demonstrate that incorporating viral genomic features in addition to host ecological traits enhanced the accuracy of potential OPV host predictions, highlighting the importance of host-virus molecular interactions in predicting potential host species. We identify hotspots for geographic regions rich with potential OPV hosts in parts of southeast Asia, equatorial Africa, and the Amazon, revealing high overlap between regions predicted to have a high number of potential OPV host species and those with the lowest smallpox vaccination coverage, indicating a heightened risk for the emergence or establishment of zoonotic OPVs. Our findings can be used to target wildlife surveillance, particularly related to concerns about mpox establishment beyond its historical range.

Introduction

Variola virus, the smallpox-causing agent belonging to the orthopoxvirus genus (OPV), has left an indelible mark on human history. This exceptionally virulent disease has triggered some of the most catastrophic outbreaks in human memory, leading to significant morbidity and mortality on a global scale. However, it also catalyzed progress in therapeutic intervention, inspiring the development of the first and highly effective vaccine. This vaccine leveraged cross-protective immunity from a closely related but less virulent OPVs, ultimately leading to the successful eradication of smallpox. In 1980, smallpox became the first human disease to be eradicated as a result of a global, coordinated vaccination effort --- and only made possible by a lack of suitable animal reservoirs to maintain variola virus outside of human populations. After the eradication of smallpox, vaccinations were largely discontinued, resulting in a worldwide reduction in immunological memory against OPVs. Nevertheless, the OPV genus is a diverse group, with many members still circulating in animal reservoirs, facilitating periodic emergence in humans and complicating intervention efforts.

Orthopoxviruses are notable for their varied mammalian host breadth and pathogenicity, although the full range of circulation in wildlife is unknown (1). The diversity in host breath has been attributed to variations in the large repertoire of accessory genes found across poxvirus species including many genes associated with the inhibition of host innate immune responses (3–5). The evolutionary dynamics of OPVs is marked by gains and losses among accessory genes, with genome reduction thought to contribute to host specialization (1). For example, the modified vaccinia virus Ankara (MVA), a third-generation smallpox vaccine, lost considerable genetic information (roughly 15% of the ~200kb genome), including many genes used by OPVs to regulate the mammalian host environment, resulting in a restricted host range (6). The viral genome of cetaceanpox viruses, a putative sister clade to OPVs, encodes for roughly half the number of proteins found in other poxviruses (7), and these viruses are thought to have highly restricted host ranges. Other OPVs, including mpox virus (formerly known as monkeypox virus) and cowpox virus, are capable of infecting a broad range of hosts, increasing the likelihood of recurrent spillover events into the human population (1). The unprecedented recent global spread of mpox virus has raised concerns about human-to-animal spillback to susceptible hosts outside its historical range, as has been observed with SARS-CoV-2 in white-tailed deer populations (8). Furthermore, the recent emergence of the novel alaskapox virus in humans from an unknown animal reservoir suggests a high likelihood of unidentified OPVs in nature with zoonotic potential (9).

Given this complex landscape of host ranges and zoonotic potential, identifying possible animal reservoir hosts of OPVs is critical for anticipating spillover events. However, identifying reservoirs through field sampling is a resource-intensive endeavor (10). To target these efforts, statistical models can be used to predict undiscovered hosts of zoonotic viruses. These statistical models leverage intrinsic traits of the virus (11), host (4, 5), or both to predict “missing links” in the host-virus network (a many-to-many predictive problem, often called *link prediction*), or can focus more specifically on host susceptibility to a specific pathogen or pathogen group of interest *(*a many-to-one predictive problem, which we call *host prediction*). This latter approach is more easily implemented, given easy access to datasets that capture host ecological, geographic, morphological, and phylogenetic features (13), and a lack of comparable standardized datasets of “viral traits.”

These types of modeling studies may be undermined by their reliance on host ecological and phylogenetic features, and conversely, the lack of predictors that capture host-virus compatibility at molecular scales. For example, several such modeling efforts have indicated a high probability of compatibility between domestic pigs (*Sus scrofa*) and SARS-CoV-2 (13, 15, 16), which was further supported by *in vitro* susceptibility of pig-derived cell lines (17). However, *in vivo* challenge studies demonstrated that SARS-CoV-2 fails to infect domestic pigs (18, 19), suggesting a post-entry incompatibility that is poorly understood or captured in current predictive models. Similarly, the Egyptian fruit bat, *Rousettus aegyptiacus*, was predicted to be a compatible host for nipah virus using host trait data (20), but failed to support nipah virus replication with experimental challenge (21). Furthermore,

viruses evolve over time, changing host breadth or transmission potential, as illustrated by the expanded host range of the SARS-CoV-2 Omicron variant (22). As the evolution of viruses is shaped by the host environment, from cell entry to evading host innate and adaptive immune response, viral genomes encode valuable information for predicting their host range but have seldom been included in predictive modeling of host-virus associations.

Here, we developed a trait-based approach using boosted regression trees (BRTs), a machine learning algorithm commonly used in ecological and evolutionary research, and integrated both host ecological traits and viral genomic features to predict mammal-OPV associations. We aggregate candidate host genera with evidence of exposure to OPV species (based on molecular detection of viral genetic material) or that are likely susceptible to infection (based on isolation of an OPV from a host). We then extend this framework to develop a model of host-OPV associations and predict the probability of a specific link between a single OPV and a single host genus, or *link prediction*. Unlike a host trait model, which predicts positivity for a virus or group of viruses, our link prediction model predicts compatible host-virus pairs by including features critical to host-virus interactions on a molecular scale. Such predictive modelling approach can be applied to optimize targeted sampling and provide quantitative insights into the hypothesized role of viral genomic variation in shaping OPV host range.

Results

Host exposure model

Host trait BRT models that aimed at predicting the host range of OPVs using either known host exposure to OPVs (based on PCR) or known susceptible hosts of OPVs (based on virus isolation) had moderate accuracy (Figure 1; Supplementary Table 1). Although the model trained on virus isolation data had higher sensitivity and greater specificity than those trained on PCR data (Figure 1), we observed no significant difference in model accuracy (t = 1.01, p = 0.314, Table S1), resulting in a small standardized effect size (Cohen's d = 0.14, Table S1). Likewise, the predicted probabilities of OPV positivity were significantly correlated between the two evidence type models (Spearman ρ = 0.516, p<0.001), with predictions of both models also exhibiting strong phylogenetic signal (Pagel's λ = 0.80 and 0.90, respectively). Phylogenetic factorization identified some overlapping taxonomic patterns in the predictions of both host trait models, including a higher mean probability of genera from the family Felidae (n = 17 species) to host OPVs, and a lower probability of the orders Lagomorpha and Rodentia (n = 514) as well as a subclade of the rodent suborder Hystricomorpha (n = 74)(Figure 2; Supplementary Figure 1; Supplementary Tables 2 and 3).

Link prediction model

In contrast, the link prediction model trained on host traits and viral features (to predict the existence of a host-virus link) classified compatible host-virus pairs with higher accuracy, higher specificity, and higher, moderate sensitivity (Figure 1; Supplementary Table 1). To explore the robustness of the link prediction model, we tried excluding host associations with vaccinia virus, given the artificial selection of viral genes implicated in virulence, replication capacity, and host range experienced by multiple passages *in vitro*. Excluding links with vaccinia virus significantly increased sensitivity (t = -10.85, p < 0.001) and specificity (t = 6.29, p < 0.001) and lowered overall model accuracy (t = 10.33, p < 0.001), albeit marginally (AUC = 0.96 vs. 0.95) (Figure 1; Supplementary Table 1).

Similar to host trait models, link predicted probabilities, represented as the mean probability of hosting any OPV per host genera, displayed strong phylogenetic signal (Pagel's λ = 0.88), and phylogenetic factorization also predicted Felidae (n = 17 tips) more likely to host OPV and Lagomorpha and Rodentia (n = 514) less likely to host OPV (Figure 2 and Supplementary Table 4). While some overlapping taxa of the order Perissodactyla and the cetacean suborders Mysticeti and Odontoceti were predicted to host OPVs by both the link prediction and host trait models, the link prediction model uniquely predicted higher probabilities of hosting an OPV among elephants (Proboscidea) and South American sloths and anteaters in the order Pilosa (n = 11; Supplementary Table 4).

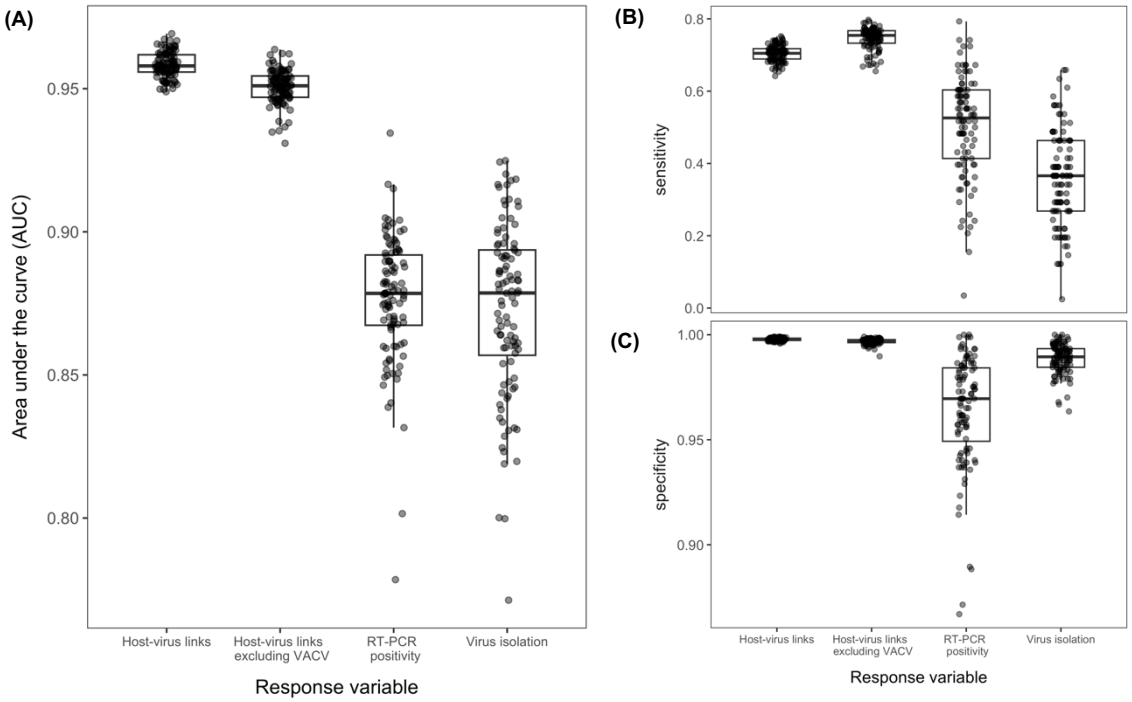

Figure 1. Performance of host-trait and link prediction BRT models as measured by the (A) areas under the receiver operating characteristic curve (AUC), (B) sensitivity and (C) specificity. Link prediction models were trained on known host-virus links vs. host-virus links excluding vaccinia virus as the response across 100 random splits of training (70%) and test (30%) data. Host-trait models were trained on RT-PCR positivity versus virus isolation data as the response across 100 random splits of training (70%) and test (30%) data.

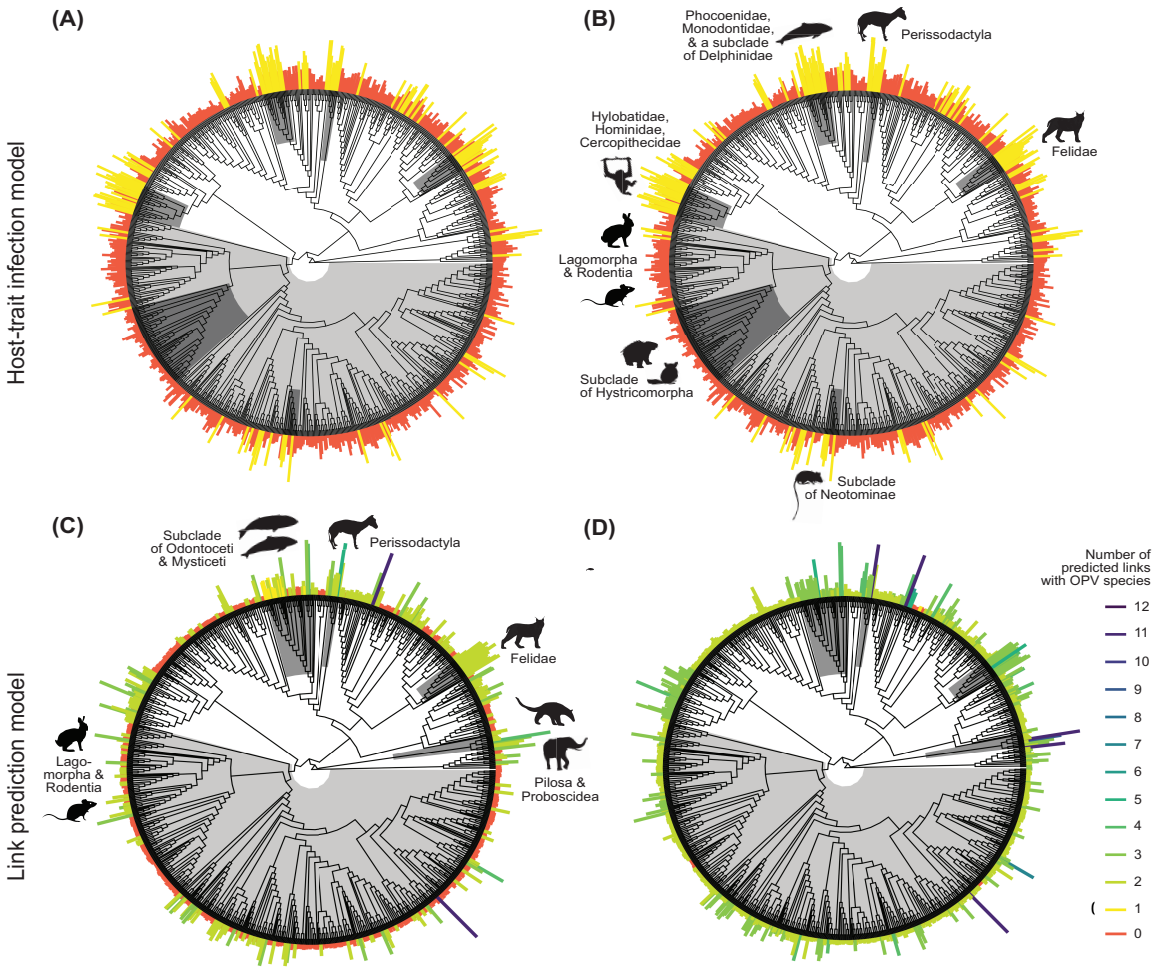

Figure 2. Predicted probabilities of orthopoxvirus positivity and taxonomic patterns in predictions identified through phylogenetic factorization for (A, B) the host exposure model and (C, D) the link prediction model. Bars/segments are scaled by probabilities and colored according to binary classification of predicted host genera assuming **(A, C)** an 80% sensitivity threshold versus **(B, D)** a threshold maximizing the sum of sensitivity and specificity. Predicted probabilities for the link prediction model, which predicted multiple links per host genera, are represented as the mean probability of hosting any OPV per host genera. Clades identified through phylogenetic factorization with significantly different mean predictions are shaded in grey with corresponding labels and representative mammal silhouettes.

Optimizing classification

To address the problem of class imbalance (i.e., the unbalanced representation of hosts and non-hosts or links and non-links in our training datasets), we explored various optimal thresholding methods for generating binary host classifications from the predictions of our host trait and link prediction models. Unsurprisingly, the number of predicted mammalian hosts for OPVs was highly sensitive to the classification threshold, particularly for the link prediction model (Figure 2 and Supplementary Table 5). Applying an 80% sensitivity threshold ($t_{ReqSens80}$ = 0.38) versus a 90% sensitivity threshold ($t_{ReqSens90}$ = 0.28) to host trait model predictions based on PCR data resulted in an approximate two-fold increase in the number of predicted host genera (from n = 121 to 249) (Supplementary Table 5). The same thresholds selection method applied to link predictions ($t_{ReqSens80}$ = 0.14 vs. $t_{ReqSens90}$ = 0.05) resulted in a

3.5-fold increase in the number of predicted host genera (from n = 470 using 80% sensitivity to n= 1,662 when using 90% sensitivity), which had a similar effect to maximizing the sum of sensitivity and specificity ($t_{MaxSensSpec}$ = 0.06; n = 1,395) (Supplementary Table 5).

Threshold selection also led to changes in the taxonomic composition of previously unobserved, predicted mammal genera when grouped by mammal order. For the link prediction model, increasing the required sensitivity from 80% to 90% (i.e., lowering the threshold), led to an increased representation of rodents (from 33% to 42%) and a decreased representation of carnivores (from 23% of 17%), while the percentage of unobserved predicted genera from other mammal orders remained the same: e.g., Cetartiodactyla (16%), primates (10-11%), and Eulipotyphla (7-8%).

Feature importance in host trait and link prediction models

We observed some overlap in the mammal traits identified as predictive of OPV positivity and compatible host-virus pairs. Across both host trait models and the link prediction model, PubMed citations (i.e., an indicator of sampling effort) and dispersal potential (i.e., the distance an animal can travel between its place of birth and its place of reproduction) were consistently important predictors (Figure 3; Supplementary Figure 2-4; Supplementary Tables 6 and 7). To determine whether citation counts, as a proxy for sampling effort, confounds the relationship between host trait profiles and OPV positivity, we conducted a secondary set of BRTs that modeled citation counts as a Poisson response, and found that host traits did not predict study effort based on its classification performance (mean AUC = 0.5, which is not different from chance).

In our link prediction model trained on host and viral traits, viral features were represented by ten principal components (PCs) derived from a principal component analysis (PCA; Supporting Methods). These components captured the majority (70%) of the variance in the presence/absence of viral accessory genes (Supplementary Figure 5). Both models based solely on host traits ranked life history traits highly (Figure 3A, Figure 3B). In contrast, rankings of relative feature importance for the link prediction model identified a multitude of PC variables as important predictors (Figure 3C). Further analysis of the influential PCs and their loadings, which indicate the relative contribution of individual genes to each PC, found potential patterns in the functional roles of genes with high positive or negative loadings (Supplementary Data S2-S5). For instance, genes with the greatest contributions to PC4 and PC3 largely encoded non-essential membrane proteins and proteins that target innate cytokines/chemokine and cell death inhibitors (Figure 3D; Supplementary Figure 6). Notably, for PC1, genes with the greatest positive loadings were predominately associated with cowpox virus, while genes with the greatest negative loadings were primarily associated with cowpox virus and mpox virus (Supplementary Data S2-S5), indicating clustering among specific OPV species (Supplementary Figure 7) that may not have been fully captured by the accessory genes included in the PCA.

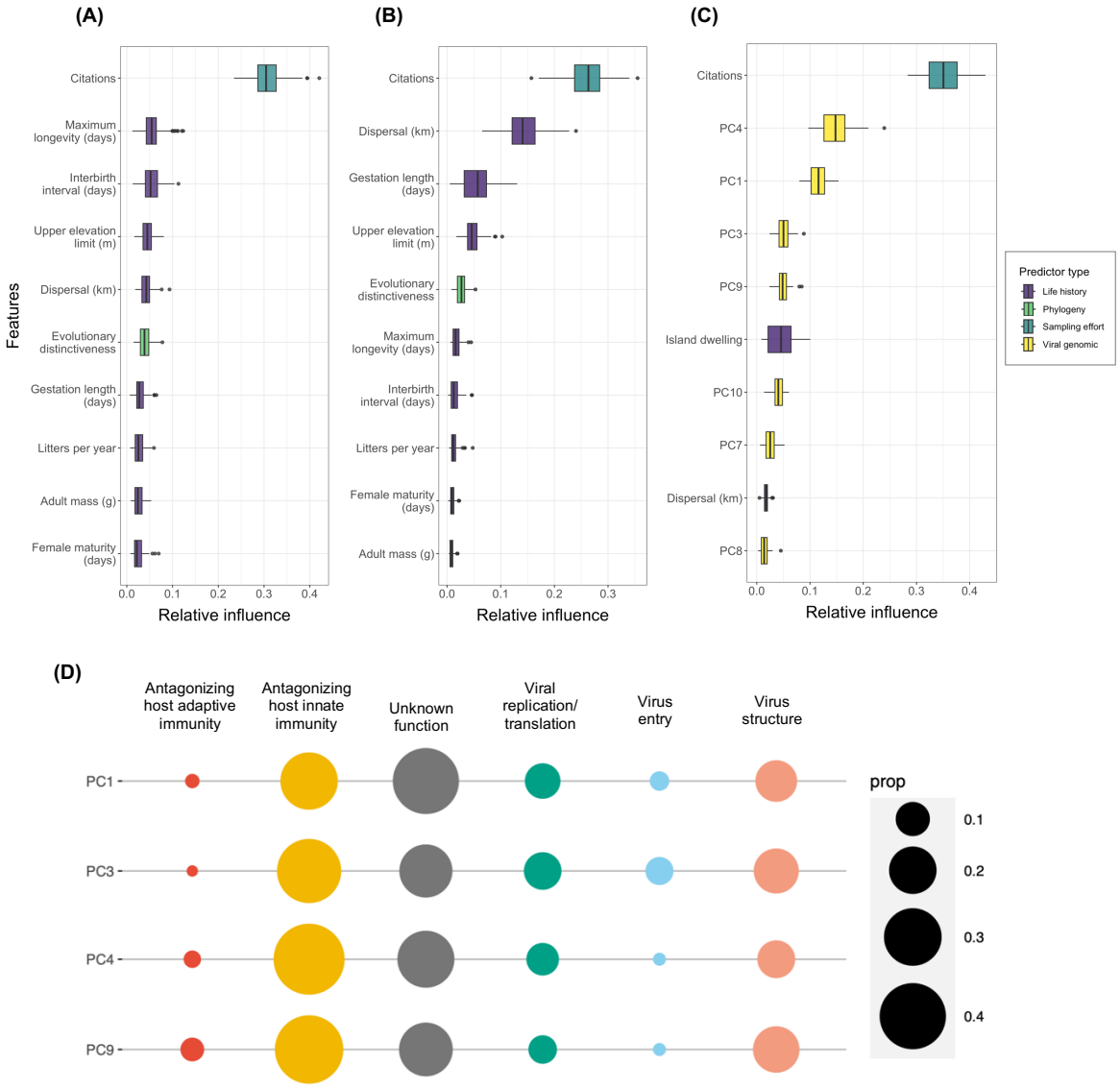

Figure 3. Relative influence of model features ranked. The most important predictor variables in BRT analysis are shown for **(A)** models trained on host traits using RT-PCR data versus **(B)** virus isolation data as response variables and **(C)** models trained on a combination of host-virus traits using link prediction (i.e., explicit pairing of host-virus links as the response) including the principal components (PCs) of viral accessory gene data. Each horizontal bar plots the variability in the relative influence of the predictor variables as measured across 100 random partitions of training (70%) and test (30%) data. Boxplots show the median and interquartile range, whiskers are the extremes, and circles are additional outliers. For the most influential viral traits based on link prediction – PC4, PC1, PC3, and PC9 – we assigned predictive functions to genes with loading values greater than 1.5 times the standard deviation in the positive range or less than 1.5 times the standard deviation in the negative range (n = 89,128, 123, and 89, respectively). The proportion of genes in each functional category is depicted in **(D)**; genes predicted to antagonize host innate immunity include those encoding Kelch-like proteins, Ankyrin proteins, and proteins involved in cell death/cell cycle regulation, chemokine/cytokine regulation, or pathogen recognition.

Distribution of potential OPV hosts

Mapping of the geographic distribution of all predicted host genera (observed and unobserved) alongside observed hosts revealed novel hotspots of predicted overlapping OPV hosts, particularly in the Eastern Himalayas, western Cameroon, and regions of Central and Eastern Africa extending from the Democratic Republic of Congo to Kenya, as well as potential novel hotspots in parts of Indonesia and Malaysia, southern East Africa, the West African coastline, the Amazon basin, and the Brazilian coastline (Figure 4A vs. Figure 4B). Our models reveal high overlap between geographic regions with the most potential OPV host species and areas where the smallest percentages of the population have been vaccinated against smallpox and are, therefore, at higher risk of zoonotic OPV emergence (Figure 4C).

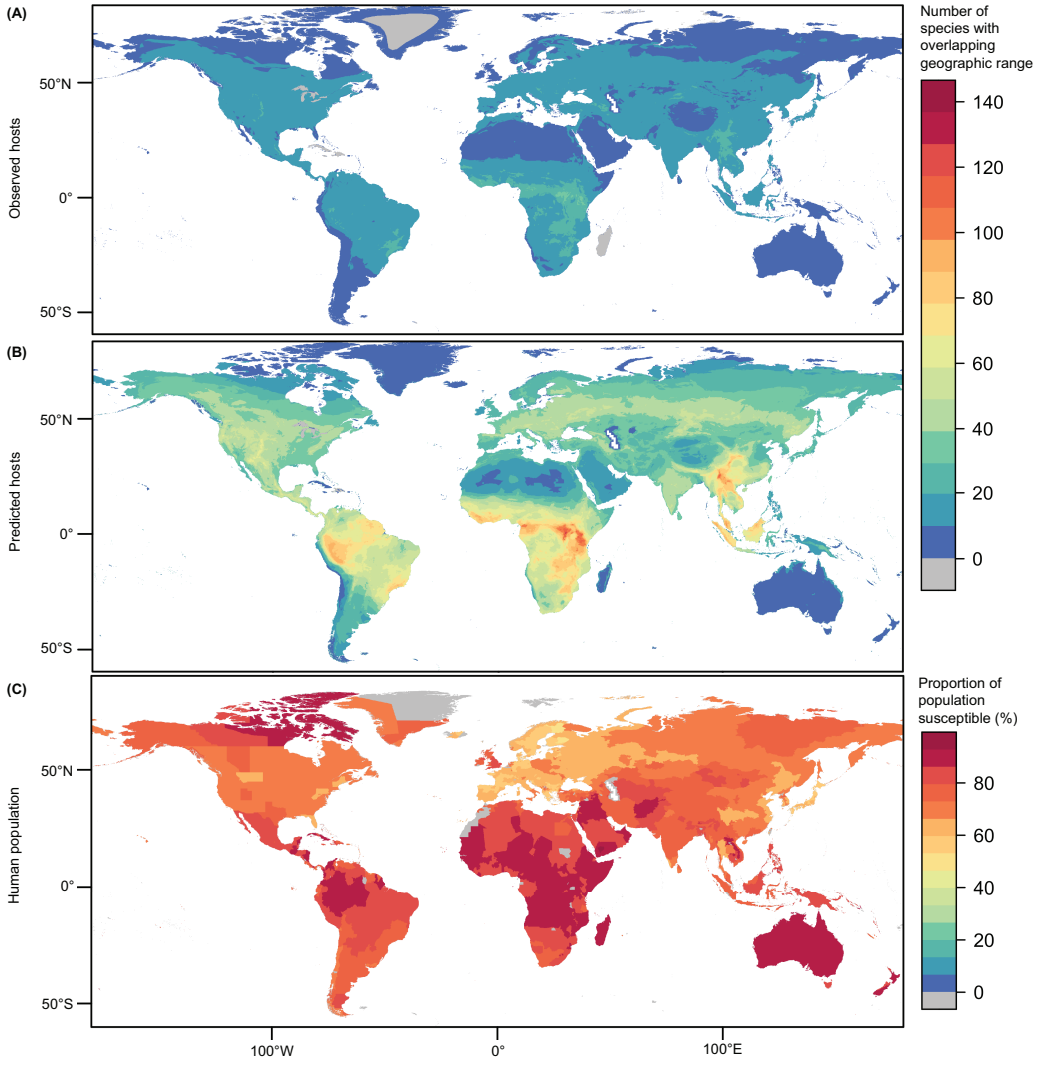

Figure 4. Geographic distribution of orthopoxvirus hosts. Host distributions are based on the IUCN Red List database of mammal geographic ranges for those species belonging to (A) observed host genera and (B) predicted (both observed and unobserved) host genera based on the results of the link prediction model and applying a 90% sensitivity threshold for host classification. The corresponding legend depicts the number of species with overlapping geographic range by color. The proportion of unvaccinated population (adapted from Taube et al. 2022) is shown in (C) whereby the legend depicts the percentage (%) of the population susceptible to orthopoxviruses.

Discussion

Machine learning models that leverage ecology to predict likely reservoirs of emerging pathogens can be remarkably accurate and can be used to guide the selection of target host species for monitoring and surveillance. However, trait selection based on host ecology alone can overlook the potential influence of genetic diversity on the ability of pathogens to infect a broad or narrow range of hosts. Here, we focused our study on OPVs, a group of closely related viruses whose genetic variability is thought to contribute to their diverse host range. Using BRTs trained on observed associations between OPVs and mammal hosts, we found that a model informed by both viral genomic features and host ecological traits predicted mammal hosts with improved accuracy compared to a model trained on host traits alone. Moreover, variables capturing viral accessory gene data were found to be highly predictive of host-virus compatibility, illustrating how integrating viral genomic traits, such as the presence or absence of virulence genes, can improve host prediction and help identify key features associated with host specificity. Relating the most influential PCs in model prediction to their original features, we classified genes with the highest loadings by their predicted functional roles, which offers new opportunities to explore the roles of accessory genes in the prediction of OPV host range.

Genomic features such as phylogenetic relatedness and codon usage bias can be informative of host prediction (11, 26, 27), but are rarely used alongside host ecological traits to inform model prediction (14). Despite distilling presence/absence data of viral accessory genes down to their principal components, our study found that BRTs trained on both host and virus traits produced similar and on average higher AUC, specificity, and sensitivity than BRTs trained on host traits alone (whether based on PCR or virus isolation data). The architecture of our model was qualitatively similar to that of Blagrove et al. (14), which used a a link prediction approach that incorporated both host and viral traits to predict potential hosts of poxviruses. Despite differences in feature selection and methodology, their model, like ours, had low to moderate sensitivity, which was attributed to an imbalance in the distribution of observed versus unobserved host-virus associations. While Blagrove et al. corrected for class imbalance using over-sampling methods, we opted to explore the effects of probability threshold selection on host predictions (14).

The problem of class imbalance is a common challenge when predicting the host range of emerging pathogens. Not only is there limited data on rare or novel hosts, but also natural variation in species abundance and host susceptibility along with sampling bias can lead to imbalance in the representation of hosts and non-hosts (i.e., compatible, and non-compatible host-virus pairs). When class imbalance is severe, applying a default threshold of 0.5 will often result in poor model performance. Thus, handling class imbalance is an important step in predictive modeling. Albeit a simple approach, tuning of the threshold parameter can have large implications on the interpretation of model predictions. For instance, our BRT PCR model trained on host traits alone estimated a 2.1 to 4.3-fold increase in the number of potential OPV hosts, assuming an 80% and 90% sensitivity threshold. Applying the same thresholds to our link prediction model trained on host and viral traits resulted in a 4.7 to 16.5-fold increase in the number of potential hosts, a likely artefact of the large proportion of pseudoabsences that construct our link prediction dataset. This severe class imbalance yielded lower predicted probabilities across BRTs making our ensemble of classifiers highly sensitive to threshold selection as compared to BRTs trained on host traits alone.

Given the wide range of predictions that can result from threshold selection, choosing the optimal thresholding method should depend on the goal of the model and the context of the predictive task. For instance, alaskapox virus is a recently emergent OPV only identified in humans thus far. Based on a more restrictive 80% sensitivity threshold, our link prediction model predicted two potentially previously unobserved host genera, *Ailurus* and *Micromys*, neither of which have a geographic range overlapping with known zoonotic events, and are thus, unlikely current reservoirs in the observed virus geographic distribution. However, applying a less restrictive threshold of 90% sensitivity resulted in the prediction of eight potential previously unobserved hosts for alaskapox virus, including the only known reservoirs of ectromelia virus, rodents in the cosmopolitan genus *Mus* (mice). Notably, genomic analysis of alaskapox virus suggests a history of recombination with ectromelia virus, including in the putative gene encoding the A-type inclusion protein (28). Deletion of the A-type inclusion protein in cowpox virus enhances viral

replication in *Mus musculus* (29), suggesting that variants or deletion mutations in this gene contribute to species specificity. We suspect this recombination site drove the indicator for *Mus* species as a potential alaskapox virus reservoir in our model. In recapitulating the ectromelia virus recombination site with alaskapox virus, our findings suggest that when sample data are sparse, applying less restrictive thresholds may be necessary to reveal informative, meaningful predictions.

Similarly, threshold selection can also impact the taxonomic representation or composition of model predictions. For instance, among known mammal genera susceptible to mpox virus, the majority are from the Rodentia (37%) and Primate (37%) orders, which is consistent with Blagrove et al. findings who predicted a similar composition of mpox virus hosts (80% from the Rodentia and Primate orders) (14). However, we show in our link prediction model, that the host composition of model predictions broaden when applying a less restrictive threshold. When assuming an 80% sensitivity threshold, 48.1% of previously unobserved mpox virus hosts were predicted to be from Rodentia, while increasing model sensitivity to 90% led to a decrease in predicted rodent genera (32%) and a substantial increase in genera from the Carnivora order (from 3.7% to 17.9%), particularly those in the Felidae (cats), Canidae (canids), Mephitidae (skunks), Mustelidae (mustelids), and Procyonidae (raccoons) families. Thus, choosing the optimal threshold for classification may have important public health implications, as it may expand the focus on species that may be overlooked when applying a more restrictive threshold. In the mpox example, this is of particular interest since its recent global spread outside endemic countries had raised concerns of the potential for transmission from humans to wildlife or spillback in nonendemic areas (32) and the establishment of new animal reservoirs.

Our study reveals a striking correspondence between regions with high potential OPV host species diversity and those where smallpox vaccination rates are lowest (Figure 4). This juxtaposition suggests a pronounced vulnerability to zoonotic OPVs within these populations. Hotspots identified as having high concentrations of postulated OPV hosts included several distinct areas within the eastern Himalayas of South Asia bordering southwest China, Central and Eastern Africa, Indonesia, and Malaysia – areas where wildlife sampling and monitoring of spillover events should focus their efforts. Comparing the geographic distribution of predicted hosts to known hosts also revealed potentially under sampled areas along the West African and Brazilian coastline; these areas also coincide with many recent human cases of reemerging zoonotic OPVs including mpox virus in Cameroon (23) and vaccinia virus in Brazil (24). Interestingly, greater geographic dispersal (distance traveled between birth site and breeding site) was not only an important predictor (Figure 3) but was also positively associated with an increased probability of hosting OPVs (Supplementary Figures 3 and 4), supporting theories of host dispersal as a key contributor to the evolution of host-pathogen dynamics (25). Specifically, as individuals move to new areas to breed, they must adapt to new environmental conditions and, in the process, are exposed to and may host more diverse pathogens.

Like all related studies, our study had several limitations, particularly with respect to incomplete datasets and model interpretability. As only complete OPV genomes were extracted from NCBI for inclusion in the link prediction model, the number of host-virus associations with accessory gene data available was limited. Moreover, missing data for host and virus traits precluded more precise predictions at the species level, and thus, we aggregated the dataset at the genera level in order to reduce the number of pseudoabsences in the models. As whole genome sequences of host and virus become more readily available, we expect future prediction models to improve in their accuracy. With respect to viral predictors, our study elected to classify accessory gene data based on presence-absence, trading simplicity over potential loss of information that could have related protein activity to host-virus compatibility. Future studies could investigate transforming accessory gene data as continuous variables based on amino acid conservation or ranked categorical variables, where accessory genes are missing, truncated and non-functional, truncated but likely functional, or intact, as previous studies have explored (33). Reducing accessory gene variables to their principal components (to avoid overfitting our model), precluded us from interpreting the effects of individual accessory genes in predicting host-OPV pairs. However, our findings could inform future iterations by training a model on the genes identified in this study as contributing highly to influential principal components, thereby directly exploring the relationship between specific viral accessory genes and host range. Furthermore, many functional roles of genes were predicted based on

sequence homology to known OPV accessory genes, particularly for vaccinia virus or mpox virus, but are not well characterized in other species. Because there is no standard for naming genes, many functions are putative and gene names can be based off other viruses, which can be misleading.

Historically, zoonotic OPVs like mpox virus were associated with small, sporadic outbreaks limited to endemic countries. However, large outbreaks in recent years including the worldwide spread of mpox virus in 2022 warrant growing concerns about the potential for mpox virus and other OPVs to establish new endemic areas. Recent evidence of animal-to-human and human-to-animal transmission of mpox virus also suggest the potential for new reservoir species to exist in traditionally non-endemic regions, underscoring the importance of predicting the host range of emerging pathogens. Adapting models to predict compatibility is also crucial to expanding the growing toolkit of analytical models used to predict host-pathogen interactions. Here, we demonstrate how trait-based machine learning models can be trained on both mammal traits of potential hosts and the genomic features of OPVs to improve the accuracy of host prediction and gain a more comprehensive understanding of the factors that influence host-virus compatibility. By exploring the trade-off between sensitivity and specificity and classification threshold selection, we show the importance of aligning models to meet the specific needs and objectives of host prediction for greater practical application.

Materials and Methods

We used OPVs as a case study to develop and validate a multi-perspective modeling framework for the prediction of host-virus associations. We generated two types of trait-based models to compare model performance and predictions: (1) a host trait model trained separately on two evidence levels for host capacity (i.e., host exposure via PCR positivity and host susceptibility via virus isolation) using host ecological traits alone; and (2) a link prediction model trained on explicit pairing of host-virus associations using a combination of host and virus features. The former predicted host positivity for OPVs, a group of viruses, while the latter predicted the existence of a host-virus link. Our modeling targets were to identify taxonomic and phylogenetic patterns in the predicted reservoir hosts of OPVs along with their corresponding geographic distribution, rank influential host and viral features, and investigate the functional roles of accessory genes that contributed substantially to link prediction.

Datasets

For the host trait model, we obtained known host positivity data from the Global Virome in One Network (VIRION), an open database of host-virus interactions drawn from scientific literature and online databases (34). We included only host positivity detected via PCR (molecular detection) or virus isolation and excluded detections with variola virus, as the virus was eradicated prior to the development of many modern diagnostics. We then collapsed host positivity data to the genus level and merged the data with the broader mammal taxonomy to obtain pseudoabsences for mammal genera with unobserved OPV associations (35). Next, we derived host predictors from COMBINE, a published mammal dataset of morphological, biogeographic, and life history traits (36). We applied summary measures to represent ecological trait data at the genus level. Specifically, we transformed all categorical variables into multiple binary variables. We then aggregated binary variables to the genus level assuming the mean to obtain the proportion of species in a genus having the variable outcome/trait, and we aggregated continuous and integer trait variables assuming the median. To account for large gaps in trait coverage, we excluded variables with zero variance or with data for less than 70% of host genera (Supplementary Figure 8). Additionally, we included binary predictors for each mammal family to represent taxonomy. We also included a variable for evolutionary distinctiveness using the *picante* package in R and a count variable of the number of scientific publications on each host genera as a measure of sampling effort using the R package *easyPubMed.* For a complete list of host predictors, see Supplementary Table 6.

For the link prediction model, we merged data from VIRION with host-OPV associations obtained from NCBI by extracting OPV genomes and their annotations (https://www.ncbi.nlm.nih.gov/)(Supplementary Dataset 1). We included viral isolates with full genome sequences from experimental laboratory studies when possible. Due to the abundance of available mpox virus and cowpox virus genomes, we included a

subset of each for all genomes with novel host associations or representing divergent clades. Multiple sequences for a particular virus-host combination were only included if an annotated genome was available and there were differences in the presence/absence of accessory genes. Again, we merged host-virus association data with the broader mammal taxonomy to accommodate out-of-sample predictions, expanding our mammal-virus network to include all combinations of mammal genera and OPV species. As predictors in our model, we incorporated the same host predictors as in the host trait model (Supplementary Table 7). Additionally, we incorporated viral predictors by identifying accessory genes from the aforementioned OPV genomes using Roary (https://sanger-pathogens.github.io/Roary/) with 80% minimum sequence identify for blastp and 95% of isolates a gene must be in to be designated as part of the core genome. After transforming the accessory gene data into a binary matrix across 981 accessory genes from 197 OPV sequences, we conducted principal components analysis (PCA) to reduce the dimensionality of our dataset and distill presence/absence variables down to their most important features. The first ten principal components, which explained roughly 70% of data variance, were then included as viral predictors in our link prediction model (Supplementary Figure 5). OPV genomic data used in the PCA was available for 12 OPV species, but only for a limited number of OPV-host pairings (e.g., Mpox virus-human, Mpox virus-dog). However, as we considered the 10 PC variables derived from the PCA as viral traits, we conducted median imputation of the PC values for the links between hosts with any of those 12 OPV species for which PC data were missing (e.g., for all unobserved Mpox virus-host we applied the median PC value for Mpox virus, for each one of the then PCs).

Lastly, for phylogenetic analysis, we used a supertree of extant mammal species trimmed to the genus level (35) (https://vertlife.org/data/mammals/). Data collation (cleaning and merging) were conducted in R version 4.2.2. Additional details regarding the cleaning of final datasets used in the host trait and link prediction models is available in the Supporting Methods section including estimation of host predictor values using summary measures, dimension reduction for generating viral predictors, and taxonomic reconciliation.

Statistical Analysis

Host trait models

Following Mull et al. (37) (https://github.com/viralemergence/hantaro), we used a trait-based model to infer the host range of OPVs, integrating only host mammal traits as predictors. This approach handles binomial virus positivity of host genera as the response variable and host trait data as predictors. It uses a machine learning algorithm to identify host features associated with OPV detection and to make predictions of the probability of OPV detection based on how similar a species' trait profile is to that of observed host species. False negative results, where a species classified as a pseudoabsence is predicted to host a virus, are used to infer potential, previously undetected or unobserved host genera.

To classify mammal genera as OPV hosts based on our matrix of predictor traits, we used boosted regression trees (BRTs). BRTs combine two validated machine learning algorithms: regression trees, which model the relationship between outcome and predictors by recursive binary splits, and boosting, an adaptive technique in which the model considers the prediction errors of previous trees when fitting subsequent trees thereby improving predictive performance and classification accuracy (38). BRTs are advantageous to our study as they can handle large ecological datasets, a binomial outcome variable, and a diversity of model predictors and the potential interactions between them. They can also fit potentially complex nonlinear relationships, produce response functions for each predictor, and are robust to missing values and outliers (39).

Using BRTs, we developed separate PCR-detection evidence models for classification of virus positivity based on PCR-detection and virus isolation. Prior to model fitting, we used the R package *rsample* to randomly partition our data into training (70%) and test (30%) sets and apply stratified sampling to ensure each dataset contained equal proportions of positive labels. Next, we tuned our model by training it on multiple combinations of hyperparameter values allowing for a maximum number of 5,000 trees, an interaction depth ranging from 2 to 4 (to control tree complexity), a shrinkage (learning) rate ranging from

0.0005 to 0.01, and a minimum number of 4 observations in terminal nodes (Supplementary Figure 9). Using the *gbm* package in R (Greenwell et al., 2020) for all model fitting, we applied five-fold cross validation and a bag/subsampling fraction of 0.5, and we assumed a Bernoulli distribution error for our binary response variable. Based on multiple model performance measures, we selected the best set of parameter values for model training, which included a learning rate of 0.01, and an interaction depth of three and the maximum number of trees set to 4500. With our tuned model for prediction, we set citation counts per genera to the mean across all mammal genera to adjust for sampling effort. We then fit replicate models to 100 randomly stratified partitions creating an ensemble of BRTs for each evidence type (PCR and virus isolation). Additionally, we constructed a third set of BRTs to determine the influence of sampling effort on the trait profiles of OPV positive genera; instead of virus positivity, we modeled the citation counts of host genera as a Poisson response variable applying the same optimal hyperparameter values except for 10,000 maximum trees.

Link prediction model

To predict the host ranges of multiple OPV species and to allow for integration of viral genomic features, we adapted the host trait model for link prediction whereby the model predicts the existence of a host-virus link (as opposed to positivity for a virus or group of viruses). While the previous host trait model predicted OPV positivity (regardless of OPV species), the link prediction model uses explicit pairing or 'links' between host genera and virus species as a response variable and, therefore, allows for the inclusion of both viral and host traits simultaneously as predictors. As in the host prediction model, undiscovered host-virus associations are identified based on false positive predictions.

Using BRTs to predict the probability of links (associations between host genera and virus species), we followed a similar workflow as the host trait model for host prediction. We checked that both test and training partitions contained similar proportions amongst the included viral taxa. We developed two ensembles of BRTs: one which included all combinations of host genera and OPV species (as described in ***Host-virus association data***), and the other which excluded any combination of Vaccinia virus. The latter was intended to explore if spontaneous, extended deletions and other modifications in viral genes implicated in virulence, replication capacity and home range of Vaccinia virus due to multiple passages in cell cultures, may introduce bias in the genetic features included in the model and affect model performance.

Model Performance and Prediction

For each model fitting of the partitioned training and test data, we obtained model performance metrics including the area under the receiver operating characteristic (ROC) curve (AUC), sensitivity and specificity using the *ROCR* and *InformationValue* packages in R and a threshold value of 0.5. For each ensemble model, we then calculated the mean and standard error of each performance metric measured across the 100 random partitions and conducted an unpaired *t*-test to compare model performance, with *p*-values adjusted for the false discovery rate using the Benjamini-Hochberg correction (40, 41).

To compare model predictions, we first calculated the Spearman correlation coefficient between the predicted probabilities of the PCR-host vs. isolation-host models as well as the link prediction model when including vs. excluding Vaccinia virus. Next, we assessed model predictions for taxonomic patterns. Because the link prediction model predicted multiple links per host genera, we calculated mean predictions per host genera to obtain a single value per tree tip (branch endpoint) for phylogenetic analysis. Using the *nlme* package in R, we estimated host phylogenetic signal based on Pagel's λ, a measure of the degree to which the propensity for virus positivity among related genera is driven by shared evolutionary history. We then identified clades with significantly different propensity for hosting OPV at various taxonomic resolutions using phylogenetic factorization, a graph-partitioning algorithm that explores the differences in measured traits between pairs of taxa while accounting for phylogenetic dependencies (42). To determine the number of phylogenetic factors (clades) that were significant, we

adjusted for the family wise error rate using Holm's sequentially rejective 5% cut-off (43). Phylogenetic factorization was implemented using the *phylofactor* package in R.

To generate binary predictions of OPV hosts and non-hosts, we transformed predicted probabilities into binary classifications using the *presenceabsence* R package. Based on the previously described ROC, we selected multiple thresholds to generate predictions based on different methods of threshold optimization including an 80% sensitivity threshold (such that 80% of observed links are detected), a 90% sensitivity threshold, and a threshold that maximizes the sum of sensitivity and specificity, otherwise known as the Youden Index (J). We then compared classifications across models and demonstrated the effects of threshold selection by plotting predicted probabilities and their binary classifications on a circular phylogenetic tree using the package *treeio* and *ggtree*. Lastly, we visualized the spatial distribution of predicted observed and unobserved mammal genera to identify and contrast geographic patterns in model predictions. Using the IUCN Red List database of mammal geographic ranges, we layered the shapefiles for thresholded species based on probabilities predicted by our models against Socioeconomic Data and Applications Center Last of the Wild's 2009 Human Footprint map, as a proxy measure for anthropogenic impact (44, 45).

Finally, we ranked model features by their mean variable importance (i.e., relative influence coefficients) and assessed similarities between models using the Spearman rank correlation coefficient, a rank-based measure of association. Viral PC predictors with high mean variable importance in our link prediction model were further analyzed by gene loading contribution to identify potential patterns in the functional roles of genes with high loadings. Specifically, we assigned predictive functions based on sequence homology to genes with loading values that were greater than 1.5 times the standard deviation in the positive range or less than 1.5 times the standard deviation in the negative range (Supplementary Dataset 2).

Limitations

Predictive models can also be limited by the machine learning algorithms they employ. As black box models, BRTs can be difficult to interpret and do not provide a complete or exact understanding of the relationships and potential interactions between predictors and the response variable. While our study employed cross-validation to assess the model's robustness, we did not perform external validation or biological validation, which requires field and laboratory studies to confirm predicted host-virus interactions. Nevertheless, our model provides guidance for empirical sampling and initial insight into the molecular signatures of host-virus compatibility. While not the aim of our study, future link prediction models informed by viral genomic data could investigate training on different types of infection evidence to gain molecular insights into host capacity.

All statistical analyses were conducted in R version 4.2.2 and Microsoft Excel version 16.69.1.

Data Availability Statement

The R code to reproduce the analyses is available from the GitHub repository: https://github.com/viralemergence/PoxHost.

Acknowledgments

This research was supported by funding to Verena (viralemergence.org) from the U.S. National Science Foundation, including NSF BII 2021909 and NSF BII 2213854. S.N.S. was partly supported by Washington State University and the US National Institute of Allergy and Infectious Disease/National Institutes of Health (NIAID/NIH) grant number U01AI151799 through the Centre for Research in Emerging Infectious Diseases—East and Central Africa (CREID-ECA).

References

1. M. G. Reynolds, S. A. J. Guagliardo, Y. J. Nakazawa, J. B. Doty, M. R. Mauldin, Understanding orthopoxvirus host range and evolution: from the enigmatic to the usual suspects. *Current Opinion in Virology* **28**, 108–115 (2018).

2. G. Sutter, B. Moss, Nonreplicating vaccinia vector efficiently expresses recombinant genes. *Proceedings of the National Academy of Sciences* **89**, 10847–10851 (1992).

3. Y. Xiang, A. White, Monkeypox virus emerges from the shadow of its more infamous cousin: family biology matters. *Emerging Microbes & Infections* **11**, 1768–1777 (2022).

4. S. N. Shchelkunov, Orthopoxvirus genes that mediate disease virulence and host tropism. *Adv Virol* **2012**, 524743 (2012).

5. S. N. Shchelkunov, An Increasing Danger of Zoonotic Orthopoxvirus Infections. *PLOS Pathogens* **9**, e1003756 (2013).

6. A. Mayr, V. Hochstein-Mintzel, H. Stickl, Abstammung, Eigenschaften und Verwendung des attenuierten Vaccinia-Stammes MVA. *Infection* **3**, 6–14 (1975).

7. T. C. S. Rodrigues, *et al.*, Genome characterization of cetaceanpox virus from a managed Indo-Pacific bottlenose dolphin (Tursiops aduncus). *Virus Research* **278**, 197861 (2020).

8. V. L. Hale, *et al.*, SARS-CoV-2 infection in free-ranging white-tailed deer. *Nature* **602**, 481–486 (2022).

9. Y. P. Springer, *et al.*, Novel Orthopoxvirus Infection in an Alaska Resident. *Clinical Infectious Diseases* **64**, 1737–1741 (2017).

10. A. S. Bernstein, *et al.*, The costs and benefits of primary prevention of zoonotic pandemics. *Science Advances* **8**, eabl4183 (2022).

11. N. Mollentze, S. A. Babayan, D. G. Streicker, Identifying and prioritizing potential human-infecting viruses from their genome sequences. *PLOS Biology* **19**, e3001390 (2021).

12. B. A. Han, J. P. Schmidt, S. E. Bowden, J. M. Drake, Rodent reservoirs of future zoonotic diseases. *Proceedings of the National Academy of Sciences* **112**, 7039–7044 (2015).

13. D. J. Becker, *et al.*, Optimising predictive models to prioritise viral discovery in zoonotic reservoirs. *The Lancet Microbe* **3**, e625–e637 (2022).

14. M. S. Blagrove, *et al.*, Monkeypox virus shows potential to infect a diverse range of native animal species across Europe, indicating high risk of becoming endemic in the region. 2022.08.13.503846 (2022).

15. I. R. Fischhoff, A. A. Castellanos, J. P. G. L. M. Rodrigues, A. Varsani, B. A. Han, Predicting the zoonotic capacity of mammals to transmit SARS-CoV-2. *Proc Biol Sci* **288**, 20211651 (2021).

16. M. Wardeh, M. Baylis, M. S. C. Blagrove, Predicting mammalian hosts in which novel coronaviruses can be generated. *Nat Commun* **12**, 780 (2021).

17. D. A. Meekins, *et al.*, Susceptibility of swine cells and domestic pigs to SARS-CoV-2. *Emerging Microbes & Infections* **9**, 2278–2288 (2020).

18. K. Schlottau, *et al.*, SARS-CoV-2 in fruit bats, ferrets, pigs, and chickens: an experimental transmission study. *The Lancet Microbe* **1**, e218–e225 (2020).

19. E. Haddock, *et al.*, Three-Week Old Pigs Are Not Susceptible to Productive Infection with SARS-COV-2. *Microorganisms* **10**, 407 (2022).

20. R. K. Plowright, *et al.*, Prioritizing surveillance of Nipah virus in India. *PLOS Neglected Tropical Diseases* **13**, e0007393 (2019).

21. S. N. Seifert, *et al.*, Rousettus aegyptiacus Bats Do Not Support Productive Nipah Virus Replication. *The Journal of Infectious Diseases* **221**, S407–S413 (2020).

22. L. Sánchez-Morales, J. M. Sánchez-Vizcaíno, M. Pérez-Sancho, L. Domínguez, S. Barroso-Arévalo, The Omicron (B.1.1.529) SARS-CoV-2 variant of concern also affects companion animals. *Frontiers in Veterinary Science* **9** (2022).

23. S. A. Babayan, R. J. Orton, D. G. Streicker, Predicting reservoir hosts and arthropod vectors from evolutionary signatures in RNA virus genomes. *Science* **362**, 577–580 (2018).

24. L. Brierley, A. Fowler, Predicting the animal hosts of coronaviruses from compositional biases of spike protein and whole genome sequences through machine learning. *PLOS Pathogens* **17**, e1009149 (2021).

25. C. M. Gigante, *et al.*, Genome of Alaskapox Virus, a Novel Orthopoxvirus Isolated from Alaska. *Viruses* **11** (2019).

26. R. J. Kastenmayer, *et al.*, Elimination of A-type inclusion formation enhances cowpox virus replication in mice: Implications for orthopoxvirus evolution. *Virology* **452–453**, 59–66 (2014).

27. D. E. Wilson, D. M. Reeder, *Mammal Species of the World: A Taxonomic and Geographic Reference* (JHU Press, 2005).

28. N. Mollentze, D. G. Streicker, Viral zoonotic risk is homogenous among taxonomic orders of mammalian and avian reservoir hosts. *Proceedings of the National Academy of Sciences* **117**, 9423–9430 (2020).

29. A. C. Fagre, *et al.*, Assessing the risk of human-to-wildlife pathogen transmission for conservation and public health. *Ecology Letters* **25**, 1534–1549 (2022).

30. , Monkeypox – Cameroon. *World Health Organization: Disease Outbreak News* (2018) (May 13, 2023).

31. J. S. de Oliveira, *et al.*, Vaccinia Virus Natural Infections in Brazil: The Good, the Bad, and the Ugly. *Viruses* **9**, 340 (2017).

32. J. Figuerola, A. J. Green, Haematozoan Parasites and Migratory Behaviour in Waterfowl. *Evolutionary Ecology* **14**, 143–153 (2000).

33. T. G. Senkevich, N. Yutin, Y. I. Wolf, E. V. Koonin, B. Moss, Ancient Gene Capture and Recent Gene Loss Shape the Evolution of Orthopoxvirus-Host Interaction Genes. *mBio* **12**, e01495-21.

34. C. J. Carlson, *et al.*, The Global Virome in One Network (VIRION): an atlas of vertebrate-virus associations. 2021.08.06.455442 (2021).

35. N. S. Upham, J. A. Esselstyn, W. Jetz, Inferring the mammal tree: Species-level sets of phylogenies for questions in ecology, evolution, and conservation. *PLOS Biology* **17**, e3000494 (2019).

36. C. D. Soria, M. Pacifici, M. Di Marco, S. M. Stephen, C. Rondinini, COMBINE: a coalesced mammal database of intrinsic and extrinsic traits. *Ecology* **102**, e03344 (2021).

37. N. Mull, C. J. Carlson, K. M. Forbes, D. J. Becker, Virus isolation data improve host predictions for New World rodent orthohantaviruses. *Journal of Animal Ecology* **91**, 1290–1302 (2022).

38. J. Elith, J. R. Leathwick, T. Hastie, A working guide to boosted regression trees. *Journal of Animal Ecology* **77**, 802–813 (2008).

39. G. De'ath, Boosted Trees for Ecological Modeling and Prediction. *Ecology* **88**, 243–251 (2007).

40. Y. Benjamini, Y. Hochberg, Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B (Methodological)* **57**, 289–300 (1995).

41. D. Yekutieli, Y. Benjamini, Resampling-based false discovery rate controlling multiple test procedures for correlated test statistics. *Journal of Statistical Planning and Inference* **82**, 171–196 (1999).

42. A. D. Washburne, *et al.*, Phylofactorization: a graph partitioning algorithm to identify phylogenetic scales of ecological data. *Ecological Monographs* **89**, e01353 (2019).

43. S. Holm, A Simple Sequentially Rejective Multiple Test Procedure. *Scandinavian Journal of Statistics* **6**, 65–70 (1979).

44. O. Venter, *et al.*, Global terrestrial Human Footprint maps for 1993 and 2009. *Sci Data* **3**, 160067 (2016).

45. O. Venter, *et al.*, Last of the Wild Project, Version 3 (LWP-3): 2009 Human Footprint, 2018 Release (2018).