Predicting the potential for zoonotic transmission and host associations for novel viruses

P. S. Pandit, S. J. Anthony, T. Goldstein, K. J. Olival, M. M. Doyle, N. R. Gardner, B. Bird, W. A. Smith, D. Wolking, K. Gilardi, C. Monagin, T. Kelly, M. Uhart, J. H. Epstein, C. Machalaba, M. K. Rostal, P. Dawson, E. Hagan, A. Sullivan, H. Li, A. A. Chmura, A. Latinne, C. Lange, T. O’Rourke, S. H. Olson, L. Keatts, A. P. Mendoza, A. Perez, C. Dejuste de Paula, D. Zimmerman, M. Valitutto, M. LeBreton, D. McIver, A. Islam, V. Duong, M. Mouiche, Z. Shi, P. Mulembakani, C. Kumakamba, M. Ali, N. Kebede, U. Tamoufe, S. Bel-Nono, A. Camara, J. Pamungkas, K. Coulibaly, E. Abu-Basha, J. Kamau, S. Silithammavong, J. Desmond, T. Hughes, E. Shiilegdamba, O. Aung, D. Karmacharya, J. Nziza, D. Ndiaye, A. Gbakima, Z. Sijali, S. Wacharapluesadee, E. Alandia Robles, B. Ssebide, G. Suzán, L. F. Aguirre, M. R. Solorio, T. N. Dhole, N. T. T. Nga, P. L. Hitchens, D. O. Joly, K. Saylor, A. Fine, S. Murray, W. Karesh, P. Daszak, J. A. K. Mazet, PREDICT Consortium & C. K. Johnson

Host-virus associations have co-evolved under ecological and evolutionary selection pressures that shape cross-species transmission and spillover to humans. Observed virus-host associations provide relevant context for newly discovered wildlife viruses to assess knowledge gaps in host-range and estimate pathways for potential human infection. Using models to predict virus-host networks, we predicted the likelihood of humans as hosts for 513 newly discovered viruses detected by large-scale wildlife surveillance at high-risk animal-human interfaces in Africa, Asia, and Latin America. Predictions indicated that novel coronaviruses are likely to infect a greater number of host species than viruses from other families. Our models further characterize novel viruses through prioritization scores and directly inform surveillance targets to identify host ranges for newly discovered viruses.
Identifying zoonotic virus emergence events at the earliest possible stage is key to mitigating outbreaks and preventing future epidemic and pandemic threats. By the time novel viruses are recognized in humans, often as a cluster of unusual cases, public health interventions to prevent or contain an epidemic face major challenges. However, determining the potential zoonotic transmission for newly discovered animal viruses, in the absence of documented human infection, is currently a major scientific challenge. New approaches are needed to evaluate and characterize the risk of zoonotic transmission of newly discovered animal viruses in the face of very limited data. Here we analyze human, domesticated animal, and wildlife surveillance and viral discovery data collected from 2009 to 2019, as part of a consortium-led One Health project aimed at strengthening pandemic threat detection capabilities in Africa, Asia, and Latin America. Surveillance efforts resulted in 944 novel monophyletic clusters of virus sequences in wildlife (referred to as novel viruses henceforth) from 18 virus families sampled at high-risk animal-human disease transmission interfaces in 34 countries. As none of these viruses have yet been identified in humans, other indices were previously established to assess potential risk, such as virus host range or plasticity, and expert opinion based on integration of ecological and molecular characteristics of viruses. We were able to quantify the risk of zoonotic transmission for 531 out of 944 novel animal viruses using data driven models to predict host-virus networks. Patterns observed across host-virus networks have been used to understand virus sharing among vertebrate species, and predict cryptic links between mammalian, and avian hosts and their viruses. Host-virus network linkages can be informed by virus traits, virus biogeography, host ecological niches, and propensity for host sharing among viruses. Precedence in viral sharing among species and ecological opportunities for spillover, as characterized by network topology, can inform propensities for newly discovered viruses that lack data. Further exploration of these networks can aid in estimating the host plasticity of viruses, an important characteristic associated with zoonotic potential. Unfortunately, systematically collected surveillance data to parameterize and validate these models have been missing. Here, we apply a network approach to gain ecological insights from viruses that have been shared among species in nature and inform potential virus-host associations and zoonotic risk of novel viruses recently discovered from wildlife.

Using data from the literature, we developed a network that included 269 known zoonotic and 307 non-zoonotic viruses infecting 885 avian and mammalian hosts. The network was used to train and validate two gradient boosting decision tree models to predict links and taxonomic orders of missing links generated by host sharing. Trained models were used to predict possible host links for 531 novel viruses due to commonalities in host sharing with known viruses and generated a predicted host-virus network (Fig. 1) formed due to the inclusion of novel viruses and their predicted linkages. We also predicted the taxonomic order of probable hosts shared as a link between two virus nodes of the network and the likelihood of the link to be humans, indicative of viruses’ predicted potential to be zoonotic.

Results and discussion

Virus-host network for known viruses ($G_c$). We developed a unipartite network with viruses as nodes and host species as edges for all species recognized as a host for viruses based on data presented in previous studies and databases, specifically, data shared by Olival et al., Pandit et al., and Johnson et al. and GenBank. In the observed network ($G_c$), viruses were represented as nodes and a link (edge) was generated if two viruses had been detected in the same host species. The observed network ($G_c$) included 576 viruses as nodes and 35,838 edges (viruses linked because of shared hosts) representing 352 vertebrate species (Fig. 1). Exploration of network characteristics of known viruses revealed differences in host sharing among virus families. The distributions of centrality measures (Fig. 2a, b, e, i) for Filoviridae, Flaviviridae, Hantaviridae, and Orthomyxoviridae families were statistically different from the mean ($p < 0.05$). Furthermore, after accounting for sampling bias for individual viruses using PubMed hits (number of search results associated with virus using a specific search term), we ran a linear regression model with node-level permutations (10,000 permutations to further characterize the distribution of viruses within virus families in the network). Viruses in families Hantaviridae, Filoviridae, Flaviviridae, and Orthomyxoviridae had a significantly higher degree ($p < 0.05$) and eigenvector centrality ($p < 0.05$), indicating more connections in the host-virus network than other representative virus families. Viruses from the Flaviviridae family also had higher betweenness centrality ($p = 0.01$) indicating more connections based on shared host species (Figs. S2–S5). Results based on distributions of centrality measures, as well as node level regression models, show similar directionality for Hantaviridae, Filoviridae, Flaviviridae, and...
Orthomyxoviridae families across multiple network topological metrics. Our findings provide further evidence for direct relationship between higher host plasticity and greater zoonotic potential.

The wildlife surveillance data consisted of tests for 99,375 animals, representing specimens from 861 species, mostly bats, rodents, primates, and other mammals (https://zenodo.org/record/5899054). To predict associations (linkages) between novel viruses with other viruses formed due to common host species, gradient boosting models were trained using network topological characteristics and families of viruses in the virus pairs to estimate: (1) whether virus pairs have a species host in common; and (2) the taxonomical order of shared hosts (Fig. 1).

**Characteristics of predicted network** ($G_{predicted}$) and newly discovered viruses. The binary model performed well in predicting the presence of links formed due to sharing of hosts between two virus nodes, and rodents, primates, and other mammals (mean positive predictive value = 0.99, sensitivity = 0.96, F-score 0.97, Fig. S6). The distribution of predicted probability for all links using the binary model showed clear bimodal distribution (Fig. S7a). The accuracy scores as a function of precision and recall indicated good model performance beyond 0.15 predicted probability for the binary model (Fig. S8). Hence, as a more conservative approach and to give more weight to the precision, we decided to use 0.7 as an optimum threshold for detecting a positive link between two virus nodes (viruses). The performance of the multilabel model varied for taxonomical orders, with a high to moderate performance for predicting taxonomical group and order of ‘humans’ and Cetartiodactyla (Figs. S7, S9). For 531 novel viruses, we identified 184,055 possible links to other viruses formed due to sharing of hosts (based on the optimum probability threshold of 0.7 identified for the binary model) generating the predicted network ($G_{predicted}$) Fig. 1, Fig. S7a). For these predicted links, between two viruses, the multiclass model was able to estimate the potential taxonomic order of the shared species for 175,113 links. For the remaining links, the model was not able to confidently predict a specific taxonomic order. Jaccard coefficient consistently showed high importance with all three importance metrics tested (gain, cover, and weight), indicating predictive ability in identifying missing links between unipartite viruses formed due to sharing of hosts (Fig. S10). Although, we suspect that these neighborhood-based topological features will be more predictive in a bipartite network setting due to more information flow between nodes.

Empirical biological networks are rarely scale-free (network with large hubs and showing a power-law distribution for degree) but a recently published study with host-host projected networks where links are represented by sharing of pathogens between hosts, has shown scale-free nature where models with power-law distributions showed the best fit for host-parasite networks. Similarly, both observed ($G_{c}$) and predicted ($G_{predicted}$) networks provided evidence that some viruses shared significantly larger numbers of hosts, creating hubs of preferential attachment and showed weak evidence of scale-free nature (loglikelihood ratio test $p > 0.05$). The degree (Fig. 2a, e, i) and betweenness centrality (Fig. 2b, i, j) distributions for predicted network ($G_{predicted}$) showed longer tails (Kolmogorov-Smirnov, $p < 0.05$) than the observed network ($G_{c}$) both at network and virus family level. Mean network degree for all virus families reduced significantly with the addition of newly discovered viruses that were predicted to have fewer links than known viruses, indicating lower host plasticity for novel viruses than known viruses or insufficient adjustment of reporting bias (Fig. S11).

**Fig. 2 Predicting missing links between virus-host communities.** Distribution shapes of degree (a) and betweenness centrality (b) for the observed and predicted network. Degree distributions for virus families in observed and predicted networks are shown in e and f. Similarly, shapes of betweenness centrality for virus families in observed and predicted networks are shown in i and j. Right panels show boxplots for novel virus families describing degree (c), betweenness centrality (d), eigenvector centrality (g), and clustering based on the predicted network formed by the binary prediction model (h).
Based on a linear regression model with node-level permutations (10,000 permutations), our adjustment for search effort (PubMed hits) was found to have no effect on the degree ($p = 0.39$, Fig. S12) and betweenness centrality ($p = 0.22$, Fig. S13), but did significantly affect the eigenvector ($p < 0.05$, Fig. S14) and clustering coefficient ($p < 0.05$, Fig. S15) of novel viruses. These results indicate that sampling and reporting efforts affect our understanding of the predilection towards certain species as illustrated by clustering in the network, but do not affect the prediction of missing host links quantified by degree centrality within the network. Many of the newly discovered viruses were mostly detected in only one species (mean = 1.32, SD ± 0.99, n = 944). Long tails of centrality distributions generated for the predicted network ($G_{predicted}$) and comparatively lower centrality measures for novel viruses, when compared with known viruses, support a tendency for newly discovered viruses to be more host-specific than previously recognized viruses, a pattern that should be further evaluated with additional sampling effort to identify the full host range for novel viruses.

Importantly, a comparison between virus families of novel viruses showed that novel coronaviruses had a higher degree ($p < 0.001$, Fig. 2c, Fig. S12), betweenness ($p = 0.02$, Fig. 2d, Fig. S13), and eigenvector ($p < 0.001$, Fig. S14) centralities in the predicted network compared to newly discovered viruses in all other virus families (Fig. 2c, d, g). In addition, the raw detection data showed significantly higher host diversity for novel coronaviruses with a mean of 2.02 (SD ± 0.70, n = 834) for other novel viruses detected in this study. This finding raises concern about the ability of novel coronaviruses to infect a greater number of species than viruses from other families. The recently emerged SARS-CoV-2 and the previously emerged SARS-CoV-1, have shown a wide host breadth16. These predictions for novel coronaviruses highlight their key ecological properties that can influence spillover into humans. Following coronaviruses, novel flaviviruses showed significantly higher betweenness centrality ($p < 0.001$). Host taxonomic order for novel viruses had no significant association with the degree centrality of the virus in the predicted network. Predicted network characteristics not only differentiate virus families based on network characteristics but also predict network characteristics that are key in understanding the ecology of a novel virus and its behavior within the network community of hosts, including the expected breadth of host species most likely to be infected by that novel virus.

Prioritizing novel viruses for further characterization. For the 531 novel viruses, we developed prioritization metrics that inform
on the ecological and evolutionary tendencies for spillover based on number of human links with known viruses predicted by the multiclass model. Novel viruses from Herpesviridae, Rhabdoviridae, Coronaviridae, Adenoviridae, Astroviridae, and Paramyxoviridae families not only showed a high median probability of sharing human links with known viruses (Fig. S16) but also were predicted to have large numbers of human links in the predicted network ($G_{predicted}$). Novel viruses of the Picobirnaviridae and Rhabdoviridae families detected here have been speculated to be hyper-parasites infecting bacteria and insects and were identified in mammalian host samples. Hence the predicted associations for these virus families should not be inferred as infection but only as detection in host samples (e.g., potentially insect viruses detected in oral swab samples from bats). Based on generalized linear mixed models, search effort (PubMed hits) was not associated with the predicted number of human links ($p = 0.24$, Table S1) nor the mean probability of sharing human links for novel viruses ($p = 0.778$, Table S2).

For a relative comparison of zoonotic risk for novel viruses, a prioritization score was developed based on the predicted probability of links being human and the number of shared human links in the predicted network for a given virus. To understand the performance of the prioritization score, we compared scores for known zoonotic and non-zoonotic viruses generated by the ensemble of both binary and multi-class models. Results indicated significantly higher prioritization scores for known zoonotic viruses (Fig. S17, $p < 0.001$) compared to known non-zoonotic viruses. Prioritization scores were derived essentially from the prediction of new/yet unobserved network links generated by the virus with another virus formed due to sharing of hosts. However, models were unable to predict new links for well-recognized viruses that have numerous hosts, such as Rabies virus and West Nile virus, and consequently resulted in a prioritization score of zero. Figure 3a–d shows the top ten and bottom five novel viruses from four virus families for relative comparison based on the prioritization score (Figs. S18–S24). PREDICT_CoV-15 found in two Phyllostomidae bats from South America (Artibeus lituratus, Sturnira lilium) scored the highest prioritization score in all novel viruses. Other top ten novel coronaviruses based on the prioritization score included viruses detected in Phyllostomidae bats (PREDICT_CoV-4, PREDICT_CoV-13, PREDICT_CoV-11, PREDICT_CoV-5). Out of these, PREDICT_CoV-11 was also detected in Mormoopidae species (Pteronotus personatus) and PREDICT_CoV-5 was found in Vespertilionidae species (Bauerus dubiaquercus) during the surveillance. These also included coronaviruses detected in Southeast Asian Pteropodidae bat species such as PREDICT_CoV-16 and PREDICT_CoV-22. PREDICT_CoV-22 was also detected in Hipposideridae bat species (Hipposideros lekaguli). PREDICT_CoV-78 detected in multiple bat and rodent species of Southeast Asia also showed a high prioritization score. These model outcomes, especially the prioritization score, provide a data-driven tool to quantify zoonotic risk for novel viruses. Even though the model is trained on numerous data points for known zoonotic and non-zoonotic viruses, individual predictions for newly discovered viruses would only require data on hosts and virus family if used within our modeling framework.

Prioritizing future surveillance. The sharing of viruses among hosts is driven by geographical overlap and synergies in ecological niches of hosts, as well as virus-specific characteristics that enable cross-species transmission. Novel viruses discovered in rodents, bats, primates, and other mammalian hosts were sampled from sites in close association with people, or at high-risk interfaces that can facilitate disease transmission in urban and rural settings. Additional surveillance across a broader taxonomic range is essential to gain insights on newly detected viruses, further inform spillover risk, and improve model predictions presented here. We used our network model and host taxonomic data in which the novel virus is first detected to prioritize host species (surveillance targets) for further surveillance of newly
discovered viruses (Supplementary Data 1). Moreover, given the recent SARS-CoV-2 pandemic we further explored surveillance targets for novel coronaviruses. Novel coronaviruses were detected in bats, rodents, birds, and primates (Fig. 4a). For novel coronaviruses, that were detected in bats, predicted surveillance targets for bat coronaviruses showed three distinct clusters (Fig. 4b). The first cluster of novel coronaviruses in bats had a higher proportion of predicted species from the Miniopterae family (Bent-winged bats) but none from Natalidae (Neotropical funnel-eared bats). Another prominent cluster prioritized all 11 chiropteran families, while the third cluster of coronaviruses showed relatively fewer host recommendations from Miniopterae bats. Representation of these surveillance targets through these clusters highlights host predilection of novel coronaviruses and indicates their preferential sharing of hosts. These clusters also support earlier results related to the scale-free nature of the predicted network ($G_{predicted}$) by creating virus hubs in the virus-host network. Cluster maps for other virus families providing evidence for future surveillance are shown in Figs. S25–S31 and Supplementary Data 2.

Grange et al. developed a tool that ranks viruses for an animal to human spillover using a risk-based approach validated by inputs from various experts from the field of virology, epidemiology, and ecology9. Our approach, on the other hand, quantifies the risk of spilloveragnostically and informs the predicted host range solely based on existing data available across the breadth of viruses and natural infections observed in free-ranging mammalian and avian hosts. Although numerous studies have been recently published that predict host-pathogen predilections, our framework quantifies the risk for viruses that have been recently discovered in animal hosts.

Network models have shown to perform well with the inclusion of ecological trait data10,17 and genome sequences18, but, with the limited data available for novel viruses, the approach provided here is an important step towards characterizing zoonotic potential for newly discovered animal viruses in the face of sparse data. These results may imply that network models are better at identifying a predictive signal when they are virus-centric (viruses as nodes and shared hosts as edges), particularly given previous host-centric work has produced mixed results when using trait-agnostic network modelling approaches17. Our network approach presents some limitations specifically for viruses that have been detected in species with limited surveillance efforts to date and are thus not part of the training data. For this reason, we were able to generate predictions for only 531 novel viruses out of 944. The remaining 413 novel viruses without predictions were detected in species that were never found positive for any virus, starkly indicating the lack of surveillance in wildlife. Further, model findings should be interpreted as associations between hosts and viruses based on the detection of viruses in samples collected from host species. These associations require further understanding around the role of hosts in the transmission ecology of viruses, especially to elaborate if hosts can serve as reservoir, amplifying, or dead-end hosts. Detection of a virus in a host species is not always correlated with that host’s ability to produce viremia for further transmission. Similarly, some of the novel viruses from Picobirnaviridae and Rhabdoviridae have been speculated to be hyperparasites and the interpretation of these detections and predicted host-associations need further investigations.

Conclusions

Novel viruses with high scores on the prioritization metrics present a strong eco-evolutionary case for further genetic and in vivo characterization to understand the risk of spillover. The scoring will help streamline in-depth in vivo characterization and develop additional hypotheses related to genetic and ecological mechanisms for cross-species transmission and zoonotic spillover. Nucleotide data associated with novel viruses presented here are short, hence the current model framework of using only host associations provides a key advantage. However, network models have shown to improve prediction capacities when nucleotide data are included as features for prediction31. These tools will improve with further surveillance and discovery of new viruses and their hosts13, ultimately informing our understanding of the mechanisms of zoonotic emergence for viruses from wildlife.

Methods

Data collection. Virus-host data was collated from various sources. Major sources included the association databases included data shared by Olival et al34, Pandit et al35, and Johnson et al33. In data provided by Olival et al (assessed September 2019), host-virus associations have been assigned a score, based on detection methods and tests that are specific and more reliable. We used associations that have been identified as the most reliable (stringent data) from Olival et al34. In addition, a query in GenBank was run to extract host reports for each GenBank submission for viruses presented in each of these three databases. Initially, for each virus name, taxonomic ID was identified using entrez.search function in biopython package. The taxonomic ID helped linked to the GenBank database, identify the ICTV lineage and associated data in PubMed22,23, NCBI TaxID closely follows the ICTV database, but some recent changes in ICTV might not always be reflected in NCBL so we manually checked names to ensure matching. This included virus genus and family information along with a standard virus name. Host data were aggregated based on the taxonomic ID and associated standard name. Finally, for each virus, a search was completed in PubMed to compile the number of hits related to the virus and their vertebrate hosts using the search terms below. The number of PubMed hits (PMH1) were used as a proxy for sampling bias33,35. The virus-host association data source is presented in supplementary code and data files (https://zenodo.org/record/5899054).

Along with the PubMed terms we also queried the nucleotide database on PubMed using the taxonomic ID to find the number of GenBank entries for these viruses (PMH2). A correlation analysis between the PMH1 and PMH2 of well-recognized known viruses showed a high correlation with each other for us to safely use GenBank hits for novel viruses during the prediction stage of the model (Fig. S32).

Development of $G_{a,c}$. a. Centrality measures of observed network ($G_{a}$)

To test if centrality measures (degree centrality, betweenness centrality, eigenvector centrality, clustering coefficient) for viral nodes in the observed network ($G_{a}$) vary significantly between viral families, we firstly used the Kolmogorov-Smirnov (KS) test. KS test is routinely used to identify distances between cumulative distribution functions of two probability distributions and is largely used to compare degree distributions of networks22,23. For each viral family, distributions of centrality measures (degree centrality, betweenness centrality, and eigenvector centrality) and clustering coefficient within the observed network ($G_{a}$) were compared with the distribution of all nodes in the network using the two-tailed KS test. Secondly, a linear regression model with virus family as a categorical variable and the number of PubMed hits as a covariate to adjust for sampling bias were fitted to understand associations of viral families with centrality measures.

$$\text{centrality measure} = \beta_0 + \beta_1 \text{PMH1} + \beta_2 \text{PMH2} + \beta_3 \text{PMH1} \times \text{PMH2}$$

After fitting the model, node-level permutations were implemented. For each random permutation, the output variable was randomly assigned to covariate values and the model was re-fitted. Finally, a $p$-value was calculated by comparing the distribution of coefficients from permutations with the original model coefficient.

Network topology feature selection. Using the observed network ($G_{a}$), multiple network topological features for all node (virus) pairs were calculated. The following are topological network features calculated. Features data type, definition and methods to calculate these features are presented in Table S3.

1. The Jaccard coefficient: a commonly used similarity metric between nodes in information retrieval, is also called an intersection of over the union for two nodes in the network. In the unipartite network generated here, it represents the proportion of common neighbor viruses from the union of neighbor viruses for two nodes. Neighbor viruses are defined as viruses with which the virus shares at least a single host.
2. Adamic/Adar (Frequency-Weighted Common Neighbors): Is the sum of inverse logarithmic degree centrality of the neighbors shared by two nodes in the network.

3. Resource allocation: Similarity score of two nodes defined by the weights of common neighbors of two nodes. Resource allocation is another measure to quantify the closeness of two nodes in the network and hence to understand the similarity of hosts they infect.

4. Preferential attachment coefficient: The mechanism of preferential attachment can be used to generate evolving scale-free networks, where the probability that a new link is connected to node $x$ is proportional to $x$. Adamic/Adar coefficients have been routinely used for generalized network prediction and have shown accuracy in identifying the importance of features for our binary model.

5. Betweenness centrality: For a node in the network betweenness centrality is the sum of the fraction of all-pairs shortest paths that pass through it. The feature that we used for training the supervised learning model was the absolute difference between betweenness centralities of two nodes. The difference between the betweenness centrality represents the difference in the sharing observed by two viruses in the pair.

6. Degree centrality: The degree centrality for a node is the number of edges it has. Between degree centralities of two nodes. Unlike the difference in betweenness centrality, the difference in degree centrality only looks at the difference in the number of observed host sharing.

7. Network clustering: All nodes were classified into community clusters using Louvain methods. A binary feature variable was generated to describe if both the nodes in the pair were from the same community or not. If both viruses are from the same community, it represents a similar host predilection than accounting for the evolutionary predication of viruses (or virus families) to infect a certain type of host.

Cross-validation and fitting generalized boosting machine (GBMs) models. A nested cross-validation was implemented for the binary model while simple cross-validation was implemented for the multiclass model (multiple output categories). The parameters of the binary model were first hyper-tuned using a cross-validated grid-search method. Values were tested using a grid search to find the best-performing model parameters that showed the highest sensitivity (recall). The parameters tested for hyper-tuning and their performance are described in the supplemental material (supplementary results and Table S5). For further cross-validation of the overall binary model, all the viruses were randomly assigned to five groups. For each fold, the viruses assigned to a group were dropped from the data, and a temporary training network ($G_t$) was constructed, assuming that this represented the current observed status of the host community. For all possible pairs in $G_t$ (both that sharing and not sharing any hosts) ten topological and viral characteristics were calculated as training features (Table S4). Categorical features were one-hot-encoded and numeric features were scaled. An XGBClassifier model with binary, logistic family was trained using the feature dataset to predict virus infections on PubMed hits for the known virus. The dataset for PubMed hits was also used to determine the optimum decision threshold for determining binary classification (Fig. S6) and a precision-recall curve was used to identify positive predictive value and sensitivity at the optimum threshold (Fig. S8).

The multiclass model was implemented in the same way, creating an observed network ($G_o$) based on species-level sharing of hosts and randomly dropping viruses to generate a training network ($G_{train}$) to train the XGBoost model. The output variables were generated based on the taxonomical orders of shared hosts. A pair of viruses can share multiple hosts, hence we trained a multioutput-multiclass model. Humans were considered an independent category of taxonomical order (label) and were given a separate label from primates. For fine-tuning the multiclass model, we started with the best performing parameters of the binary model and manually tested 5 combinations of model parameters by adjusting values of the learning rate, number of estimators, maximum depth, and minimum child weight (Supplementary code and results).

Results for feature importance are shown in supplementary results (Fig. S10).
5. The multiclass model showed higher performance for correctly classifying links as "human" hosts than other numerous avian and mammalian taxonomic orders. Hence, the multiclass model outputs were summarized into either humans or other taxonomic groups. For the novel virus, a list of known viruses with the predicted link was generated. Using the hosts of these known viruses and the taxonomic order in which the novel virus was detected, a list of most likely species was generated based on the overall frequency of the host species. For understanding the likelihood of infecting humans two factors were considered to be of importance. Firstly, the number of links where humans are predicted as shared hosts with known viruses (n) and the average model-predicted probability of those links. A representation was generated incorporating the probability and available model support in terms of number links to reflect the likelihood and compare viruses relative to each other.

To test if virus family, the taxonomic order of hosts in which novel viruses were detected, and the number of times the viruses were detected (equivalent to PubMed hits for known viruses) influence node (virus) level network centrality measures in the predicted network. \( G, \) a linear regression model was fitted with centrality measures.

\[
\text{centrality measure} = \beta_0 \text{intercept} + \beta_1 \text{ViralFamily}_{\text{average}} + \beta_2 \text{HostOrder}_{\text{average}} + \beta_3 \text{PubMed hits}
\]

For each of the random 10,000 node-level permutations, the output variable (centrality measure) was randomly assigned to covariate values and the model was re-fitted. A \( p \)-value was calculated by comparing the distributions of coefficients with the original model coefficient. These models were fitted for degree centrality, betweenness centrality, eigenvector centrality, and clustering coefficient of novel viruses in the predicted network.

Prioritization score for novel viruses. Generalized Linear Mixed Models were used to understand the association effects of virus family, taxonomic order of the host and PubMed hits on the number of predicted human links and mean probability of the predicted links. The models were fit using glm.nb in R. For relative comparison of zoonotic risk and for prioritizing novel viruses for further characterization, a prioritization metric was developed based on the predicted probability of sharing the hosts as known hosts with known viruses \( p_{\text{sharing known}} \) and the number of predicted shared human links \( n_{\text{human}} \) in the predicted network for the given virus \( G \). For each combination of different \( p_{\text{sharing known}} \) and \( n_{\text{human}} \) values, and for different appropriate relative comparisons between novel viruses. To understand the behavior of the prioritization score when predicting the zoonotic risk of novel viruses, we also compared prioritization scores of known zoonotic and non-zoonotic viruses using the Kolmogorov-Smirnov test.

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

Data reported in this paper are available at https://zenodo.org/record/5899054, https://data.usaid.gov/d?qea-hwmr and https://data.usaid.gov/d/xjvz-fnrb, https://data.usaid.gov/Global-Health-Security-in-Development-GHSD-PREDICT-Emerging-Pandemic-Threats-Project?qea-hwmr.

Code availability

Code used to develop models and generate results and figures presented in the paper are available at https://zenodo.org/record/5899054.

Received: 1 September 2021; Accepted: 4 August 2022; Published online: 19 August 2022

References

1. PREDICT Consortium. 2021. PREDICT Emerging Pandemic Threats Project. Dataset. USAID Development Data Library. https://data.usaid.gov/d?qea-hwmr.
2. Kreuder Johnson, C. et al. Spillover and pandemic properties of zoonotic viruses with high host plasticity. Sci. Rep. 5, 14830 (2015).
3. Pandit, P. S. et al. Predicting wildlife reservoirs and global vulnerability to zoonotic Flaviviruses. Nat. Commun. 9, 5425 (2018).
4. Olival, K. J. et al. Host and viral traits predict zoonotic spillover from mammals. Nature 546, 646–650 (2017).
5. Grange, Z. L. et al. Ranking the risk of animal-to-human spillover for newly discovered viruses. Proc. Natl Acad. Sci. 118, e2002324118 (2021).
6. Gomez, J. M., Nunn, C. L. & Verdu, M. Centrality in primate-parasite networks reveals the potential for the transmission of emerging infectious diseases to humans. Proc. Natl Acad. Sci. USA 110, 7738–7744 (2013).
7. Albery, G. F. et al. The science of the host-virus network. Nat. Microbiol. 6, 1483–1492 (2021).
8. Walker, J. G., Pien, M., Morgan, E. R. & Veski, P. A. Uncertain links in host-parasite networks: lessons for parasite transmission in a multi-host system. Philos. Trans. R. Soc. Lond. B Biol. Sci. 372, https://doi.org/10.1098/rstb.2016.0095 (2017).
9. Dallas, T., Park, A. W. & Drake, J. M. Predicting cryptic links in host-parasite networks. PLoS Comput. Biol. 13, e1005557 (2017).
10. Albery, G. F., Eskew, E. A., Ross, N. & Olival, K. J. Predicting the global mammalian viral sharing network using phylogeography. Nat. Commun. 11, 5126 (2020).
11. Wardhe, M., Blagrove, M. S., Sharkey, K. J. & Baylis, M. Divide-and-conquer: machine-learning integrates mammalian and viral traits with network features to predict virus-mammal associations. Nat. Commun. 12, 1–15 (2021).
12. Chen, T. & Guestrin, C. In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. 785–794 (ACM).
13. Johnson, C. K. et al. Global shifts in mammalian population trends reveal key predictors of virus spillover risk. Philos. Trans. R. Soc. Lond. B Biol. Sci. 367, 2864–2871 (2012).
14. Clark, K., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J. & Sayers, E. W. GenBank. Nucleic acids Res. 44, D67–D72 (2016).
15. Schoch, C. L. et al. NCBI Taxonomy: a comprehensive update on curation, resources and tools. Database 2020, 1–21 (2020).
16. Mossman, K. & Baker, M. L. Zoonoanthropic potential of SARS-CoV-2 and implications of reintroduction into human populations. Cell Host Microbe 29, 160–164 (2021).
17. Becker, D. J. et al. Optimising predictive models to prioritise viral discovery in zoonotic reservoirs. The Lancet Microbe 3, e525–e537 (2022).
18. Mollentze, N., Babayan, S. & Streicher, D. Identifying and prioritizing potential human-infecting viruses from their genome sequences. bioRxiv 2020.2011.202012.2013.79917, 1–25 (2021).
19. Woolhouse, M., Scott, F., Hudson, Z., Howey, R. & Chase-Topping, M. Human viruses: discovery and emergence. Philos. Trans. R. Soc. Lond. B Biol. Sci. 367, 2864–2871 (2012).
20. Baum, L. & Zhou, T. Link prediction in complex networks: A survey. Phys. A: Stat. Mech. Appl. 390, 1150–1170 (2011).
21. Barabasi, A. L. & Albert, R. Emergence of scaling in random networks. Science 286, 509–512 (1999).
22. Blondel, V. D., Guillaume, J.-L., Lambiotte, R. & Lefebvre, E. Fast unfolding of communities in large networks. J. Stat. Mech.: Theory Exp. 2008, P0008 (2008).
23. Irwin, D. M., Kocher, T. D. & Wilson, A. C. Evolution of the cytochrome b gene of mammals. J. Mol. Evol. 32, 128–144 (1991).

Author contributions

P.S.P., C.K.J., S.J.A., T.G., K.L.O. and J.A.K.M conceived of the research; P.S.P. analyzed the data; P.S.P., S.J.A., T.G., K.L.O., M.M.D., N.R.G., B.B., W.A.S., D.W., K.G., C.M., T.K., M.U., J.H.E., C.M., M.K.R., P.D., E.H., A.S., H.L., A.C.C., A.L., C.I., T.O.R, S.H.O., L.K., A.P.M., A.P., C.D. de P., D.M., M.L., D.M., A.I., Y.D., M.M., Z.M., P.M., C.K., M.A., N.K., U.T., S.B. N.A., J.C., P.C., E.A., B.K., J.S., J.D., T.H., E.S., O.A., D.K., J.N., D.N., A.G., Z.S., S.W., E.A., R.S., G.S., L.F.A., M.R.S., T.N.D., N.T., T.N., P.L.H., D.O.J., K.S., A.F., S.M., W.K., P.D., J.A.K.M., PREDICT Consortium, & C.K.J. collected data, wrote and revised the manuscript.

Funding

This work was supported by the United States Agency for International Development (USAID) Emerging Pandemic Threat PREDICT program (Cooperative Agreement nos. GHN-A-00-09-000100-00 and AID-OAA-A-14-00102). P.S.P., C.K.J., M.U., K.G., and N.R.G. are also supported by the National Institute Of Allergy And Infectious Diseases of the National Institutes of Health under Award Number 1U01AI15814. The content is solely the responsibility of the authors and does not necessarily represent the official views of USAID, National Institutes of Health, or the United States Government. We thank the governments of Bangladesh, Bolivia, Brazil, Cambodia, Cameroon, China, DR Congo, Egypt, Ethiopia, Gabon, Ghana, Guinea, India, Indonesia, Ivory Coast, Jordan,
Kenya, Lao PDR, Liberia, Malaysia, Mexico, Mongolia, Myanmar, Nepal, Peru, Republic of Congo, Rwanda, Senegal, Sierra Leone, Tanzania, Thailand, Uganda, and Vietnam for permission to conduct this study, and the field teams and collaborating laboratories that performed sample collection and testing.

Competing interests
The authors declare no competing interests.

Additional information
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s42003-022-03797-9.

Correspondence and requests for materials should be addressed to P. S. Pandit or C. K. Johnson.

Peer review information Communications Biology thanks the anonymous reviewers for their contribution to the peer review of this work. Primary Handling Editors: Pei Hao and Luke R. Grinham.

1One Health Institute, School of Veterinary Medicine, University of California, Davis, Davis, CA 95616, USA. 2Center for Infection and Immunity, Columbia University, New York, NY 10032, USA. 3EcoHealth Alliance, 520 Eighth Avenue, New York, NY 10018, USA. 4Labyrinth Global Health, Inc., 546 15th Ave NE, St Petersburg, FL 33704, USA. 5Wildlife Conservation Society, Health Program, Bronx, NY, USA. 6Wildlife Conservation Society (WCS), Peru Program, Lima, Peru. 7Global Health Program, Smithsonian’s National Zoological Park and Conservation Biology Institute, Washington, DC, USA. 8Mosaic/Global Viral Cameroon, Yaoundé, Cameroon. 9Metabiota Inc, Nanaimo, VC, Canada. 10Institut Pasteur du Cambodge, 5 Monivong BlvdPO Box 983 Phnom Penh 12201, Cambodia. 11Wuhan Institute of Virology, Chinese Academy of Sciences, Wuhan, China. 12Kinshasa School of Public Health, University of Kinshasa, Kinshasa, Democratic Republic of the Congo. 13Metabiota Inc., Kinshasa, Democratic Republic of the Congo. 14Egypt National Research Centre, 12311 Dokki, Giza, Egypt. 15Aklilu Lemma Institute of Pathobiology, Addis Ababa University, Addis Ababa, Ethiopia. 16Metabiota Cameroon Ltd, Yaoundé, Centre Region Avenue Mvog-Fouda Ada, Av 1.085, Carrefour Intendance, Yaoundé, BP 15939, Cameroon. 17Military Veterinary (Rtd.), P.O. Box CT2585 Accra, Ghana. 18Centre de Recherche en Virologie (VRV) Projet Fièvres Hemorragiques en Guinée, BP 5680 Nongo/Contéyama-Commune de Ratoma, Guinea. 19Primate Research Center, Bogor Agricultural University, Bogor 16151, Indonesia. 20Faculty of Veterinary Medicine, Bogor Agricultural University, Darmaga Campus, Bogor 16680, Indonesia. 21Department Environment and Health, Institut Pasteur de Côte d’Ivoire, PO BOX 490 Abidjan 01, Ivory Coast. 22Department of Basic Medical Veterinary Sciences, College of Veterinary Medicine, Jordan University of Science and Technology, Ar-Ramtha, Jordan. 23Molecular Biology Laboratory, Institute of Primate Research, Nairobi, Kenya. 24Department of Biochemistry, University of Nairobi, Nairobi, Kenya. 25Conservation Medicine, Sungai Buloh, Selangor, Malaysia. 26Wildlife Conservation Society (WCS), Mongolian Program, Ulaanbaatar, Mongolia. 27Center for Molecular Dynamics Nepal (CMDN), Thapathali -11, Kathmandu, Nepal. 28Regional Headquarters, Mountain Gorilla Veterinary Project, Musanze, Rwanda. 29Université Cheikh Anta Diop, BP 5005 Dakar, Sénégal. 30Metabiota, Inc. Sierra Leone, Freetown, Sierra Leone. 31Department of Veterinary Medicine and Public Health, College of Veterinary Medicine and Biomedical Sciences, Sokoine University of Agriculture, Morogoro, Tanzania. 32Thai Red Cross Emerging Infectious Diseases Clinical Center, King Chulalongkorn Memorial Hospital, Bangkok, Thailand. 33Wildlife Conservation Society (WCS), Bolivia Program, La Paz, Bolivia. 34Facultad de Medicina Veterinaria y Zootecnia, Universidad Nacional Autónoma de México, México City 04510, Mexico. 35Centro de Biodiversidad y Genética, Universidad Mayor de San Simón, Cochabamba, Bolivia. 36Laboratorio de Epidemiología e Geoprocessamiento (EpiGeo), Instituto de Medicina Veterinaria (IMV) Universidad Federal do Pará (UFPA), BR-316 Km 31, Castanhal, Pará 69746-360, Brazil. 37Department of Microbiology, Sanjay Gandhi Post Graduate Institute of Medical Sciences, Lucknow, Uttar Pradesh, India. 38Wildlife Conservation Society (WCS), Vietnam Program, Hanoi, Vietnam. 39Melbourne Veterinary School, Faculty of Veterinary and Agricultural Sciences, University of Melbourne, Werribee, VIC 3030, Australia. 40Nyati Health Consulting, 2175 Dodds Road, Nanaimo V9X0A4, Canada. 41These authors contributed equally: S. J. Anthony, T. Goldstein. 42*Correspondence: T. Goldstein. 43*email: tspandit@ucdavis.edu; ckjohnson@ucdavis.edu

PREDICT Consortium
P.S. Pandit, S. J. Anthony, T. Goldstein, K. Gilardi, C. Monagin, T. Kelly, M. Uhart, J. H. Epstein, M. K. Rostal, P. Dawson, H. Hagan, A. Sullivan, H. Li, A. A. Chmura, A. Latinne, C. Lange, T. O’Rourke, S. H. Olson, L. Keatts, A. P. Mendoza, A. Perez, C. Dejuste de Paula, D. Zimmerman, M. Valittuto, M. LeBreton, D. McIver, A. Islam, V. Duong, M. Mouiche, Z. Shi, P. Mulembakani, C. Kukammba, M. Ali, N. Kebede, U. Tamoufe, S. Bel-Nono, A. Camara, A. Pamungkas, K. Coulibaly, E. Abu-Basha, J. Kamau, S. Silithammavong, J. Desmond, T. Hughes.
A full list of members and their affiliations appears in the Supplementary Information.
Terms and Conditions

Springer Nature journal content, brought to you courtesy of Springer Nature Customer Service Center GmbH ("Springer Nature"). Springer Nature supports a reasonable amount of sharing of research papers by authors, subscribers and authorised users ("Users"), for small-scale personal, non-commercial use provided that all copyright, trade and service marks and other proprietary notices are maintained. By accessing, sharing, receiving or otherwise using the Springer Nature journal content you agree to these terms of use ("Terms"). For these purposes, Springer Nature considers academic use (by researchers and students) to be non-commercial.

These Terms are supplementary and will apply in addition to any applicable website terms and conditions, a relevant site licence or a personal subscription. These Terms will prevail over any conflict or ambiguity with regards to the relevant terms, a site licence or a personal subscription (to the extent of the conflict or ambiguity only). For Creative Commons-licensed articles, the terms of the Creative Commons license used will apply.

We collect and use personal data to provide access to the Springer Nature journal content. We may also use these personal data internally within ResearchGate and Springer Nature and as agreed share it, in an anonymised way, for purposes of tracking, analysis and reporting. We will not otherwise disclose your personal data outside the ResearchGate or the Springer Nature group of companies unless we have your permission as detailed in the Privacy Policy.

While Users may use the Springer Nature journal content for small scale, personal non-commercial use, it is important to note that Users may not:

1. use such content for the purpose of providing other users with access on a regular or large scale basis or as a means to circumvent access control;
2. use such content where to do so would be considered a criminal or statutory offence in any jurisdiction, or gives rise to civil liability, or is otherwise unlawful;
3. falsely or misleadingly imply or suggest endorsement, approval, sponsorship, or association unless explicitly agreed to by Springer Nature in writing;
4. use bots or other automated methods to access the content or redirect messages;
5. override any security feature or exclusionary protocol; or
6. share the content in order to create substitute for Springer Nature products or services or a systematic database of Springer Nature journal content.

In line with the restriction against commercial use, Springer Nature does not permit the creation of a product or service that creates revenue, royalties, rent or income from our content or its inclusion as part of a paid for service or for other commercial gain. Springer Nature journal content cannot be used for inter-library loans and librarians may not upload Springer Nature journal content on a large scale into their, or any other, institutional repository.

These terms of use are reviewed regularly and may be amended at any time. Springer Nature is not obligated to publish any information or content on this website and may remove it or features or functionality at our sole discretion, at any time with or without notice. Springer Nature may revoke this licence to you at any time and remove access to any copies of the Springer Nature journal content which have been saved.

To the fullest extent permitted by law, Springer Nature makes no warranties, representations or guarantees to Users, either express or implied with respect to the Springer nature journal content and all parties disclaim and waive any implied warranties or warranties imposed by law, including merchantability or fitness for any particular purpose.

Please note that these rights do not automatically extend to content, data or other material published by Springer Nature that may be licensed from third parties.

If you would like to use or distribute our Springer Nature journal content to a wider audience or on a regular basis or in any other manner not expressly permitted by these Terms, please contact Springer Nature at

onlineservice@springernature.com