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

Pranav Pandit (pspandit@ucdavis.edu)
University of California Davis https://orcid.org/0000-0001-7649-0649

Simon Anthony
Columbia University

Tracey Goldstein
University of California, Davis https://orcid.org/0000-0002-1672-7410

Kevin Olival
EcoHealth Alliance https://orcid.org/0000-0003-3211-1875

Megan Doyle
University of California Davis

Nicole Gardner
University of California, Davis https://orcid.org/0000-0002-4046-9685

Brian Bird
University of California Davis

Woutrina Smith
University of California Davis

David Wolking
University of California Davis

Kristen Gilardi
University of California Davis

Corina Monagin
University of California Davis

Terra Kelly
University of California Davis

Marcela Uhart
University of California, Davis https://orcid.org/0000-0003-3525-541X

Jonathan Epstein
EcoHealth Alliance https://orcid.org/0000-0002-1373-9301

Catherine Machalaba
EcoHealth Alliance

Melinda Rostal
EcoHealth Alliance

Patrick Dawson

EcoHealth Alliance

Emily Hagan
EcoHealth Alliance

Ava Sullivan
EcoHealth Alliance

Hongying Li
Ecohealth Alliance

Aleksei Chmura
EcoHealth Alliance

Alice Latinne
EcoHealth Alliance

Christian Lange
Labyrinth Global Health, Inc

Tammie O'Rourke
Labyrinth Global Health, Inc

Sarah Olson
Wildlife Conservation Society https://orcid.org/0000-0002-8484-9006

Lucy Keatts
University of California Davis

A. Patricia Mendoza
Wildlife Conservation Society https://orcid.org/0000-0002-8631-0119

Alberto Perez
Wildlife Conservation Society

Catia Dejuste de Paula
Wildlife Conservation Society

Dawn Zimmerman
Smithsonian Conservation Biology Institute

Marc Valitutto
Smithsonian's National Zoological Park and Conservation Biology Institute

Matthew LeBreton
Mosaic (Environment, Health, Data, Technology)

David McIver
Metabiota Inc.

Ariful Islam
EcoHealth Alliance

Veasna Duong
Institut Pasteur du Cambodge https://orcid.org/0000-0003-0353-1678

Moctar Mouiche
Mosaic/Global Viral Cameroon

Zheng-Li Shi
Wuhan Institute of Virology https://orcid.org/0000-0001-8089-163X
Prime Mulembakani
University of Kinshasa
Charles Kumakamba
Metabiota Inc.
Mohamed Ali
Public Health Initiative, Cairo, Egypt.
Nigatu Kebede
Addis Ababa University
Ubald Tamoufe
Metabiota Inc.
Samuel Bel-Nono
Military Veterinarian (Rtd.)
Alpha Camara
Centre de Recherche en Virologie (VRV)
Joko Pamungkas
Julien Kalpy Coulibaly
Institut Pasteur de Côte d'Ivoire
Ehab Abu-Basha
Jordan University of Science and Technology
Joseph Kamau
University of Nairobi
Soubanh Silithammavong
Metabiota Inc,
James Desmond
EcoHealth Alliance
Tom Hughes
Conservation Medicine https://orcid.org/0000-0002-5713-9738
Enkhtuvshin Shiilegdamba
Wildlife Conservation Society, Mongolia Program
Ohnmar Aung
Smithsonian's National Zoological Park and Conservation Biology Institute
Dibesh Karmacharya
Center for Molecular Dynamics
Julius Nziza
Mountain Gorilla Veterinary Project
Daouda Ndiaye
Université Cheikh Anta Diop

Aiah Gbakima
Metabiota, Inc
Zikankuba Sijali
Sokoine University of Agriculture
Supaporn Wacharapluesadee
Chulalongkorn University
Erika Alandia Robles
Wildlife Conservation Society (WCS) – Bolivia Program
Benard Ssebide
Mountain Gorilla Veterinary Project
Gerardo Suzán
Universidad Nacional Autónoma de México
Luis Aguirre
Universidad Mayor de San Simón https://orcid.org/0000-0002-4194-1523
Monica Solorio
Universidade Federal do Pará
Tapan Dhole
Sanjay Gandhi Post Graduate Institute of Medical Sciences
Peta Hitchens
Swedish University of Agricultural Sciences https://orcid.org/0000-0002-7528-7056
Damien Joly
Metabiota Inc
Karen Saylors
Labyrinth Global Health, Inc
Amanda Fine
Wildlife Conservation Society
Suzan Murray
Smithsonian's National Zoological Park and Conservation Biology Institute
William Karesh
EcoHealth Alliance
Peter Daszak
EcoHealth Alliance https://orcid.org/0000-0002-2046-5695
Jonna Mazet
UC Davis
PREDICT Consortium
University of California Davis
Christine Johnson
Karen C. Drayer Wildlife Health Center and EpiCenter for Disease Dynamics, One Health Institute, University of California Davis School of Veterinary Medicine

Article

Keywords:

Posted Date: January 25th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-846253/v1

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

Authors: P. S. Pandit[1*], S.J. Anthony[2**], T. Goldstein[1**], K. J. Olival[3], M. M. Doyle[1], N. R. Gardner[1], B. Bird[1], W. Smith[1], D. Wolking[1], K. Gilardi[1], C. Monagin[1], T. Kelly[1], M. Uhart[1], J. H. Epstein[3], C. Machalaba[3], M. K. Rostal[3], P. Dawson[3], E. Hagan[3], A. Sullivan[3], H. Li[3], A. A. Chmura[3], A. Latinne[3], C. Lange[4], T. O'Rourke[4], S. Olson[5], L. Keatts[1], P. Mendoza[5], A. Perez[5], C. Dejuste de Paula[5], D. Zimmerman[6], M. Valitutto[6], M. LeBreton[7], D. McIver[8], A. Islam[3], V. Duong[9], M. Mouiche[7], Z. Shi[10], P. Mulembakani[11], C. Kumakamba[12], M. Ali[13], N. Kebede[14], U. Tamoufe[15], S. Bel-Nono[16], A. Camara[17], J. Pamungkas[18, 19], K. Coulibaly[20], E. Abu-Basha[21], J. Kamau[22], S. Silithammavong[8], J. Desmond[3], T. Hughes[3,23], E. Shiilegdamba[24], O. Aung[6], D. Karmacharya[25], J. Nziza[26], D. Ndiaye[27], A. Gbakima[28], Z. Sijali[29], S. Wacharapluesadee[30], E. Alandia Robles[31],B. Ssebide[26], G. Suzán[32], L. F. Aguirre[33], M. R. Solorio[34], T. N. Dhole[35], P. L. Hitchens[36], D. O. Joly[37], K. Saylors[4], A. Fine[6], S. Murray[7], W. Karesh[3], P. Daszak[3], J. A. K. Mazet[1], PREDICT Consortium, & C. K. Johnson[1*]

*Corresponding author: P.S. Pandit, pspandit@ucdavis.edu and C. K. Johnson, ckjohnson@ucdavis.edu

Affiliations:

[1]One Health Institute, School of Veterinary Medicine, University of California, Davis, Davis, CA, 95616, USA.

[2] Center for Infection and Immunity, Columbia University, New York, NY, 10032, USA.

[3] EcoHealth Alliance, 520 Eighth Avenue, New York, NY, 10018, USA.

[4]Labyrinth Global Health, Inc., 546 15th Ave NE, St Petersburg, FL 33704, USA.

[5]Wildlife Conservation Society, Bronx, NY, 10460, USA.

[6]Global Health Program, Smithsonian's National Zoological Park and Conservation Biology Institute, Washington, District of Columbia, United States of America

[7]Mosaic/Global Viral Cameroon, Yaoundé, Cameroon

[8]Metabiota Inc, Nanaimo, Canada

[9]Institut Pasteur du Cambodge, 5 Monivong Blvd, PO Box 983, Phnom Penh, 12201, Cambodia.

[10]CAS Key Laboratory of Special Pathogens, Wuhan Institute of Virology, Center for Biosafety Mega-Science, Chinese Academy of Sciences, Wuhan, China.

[11]Kinshasa School of Public Health, University of Kinshasa, Kinshasa, Democratic Republic of the Congo.

[12]Metabiota Inc., Kinshasa, Democratic Republic of the Congo

[13]Egypt National Research Centre, 12311 Dokki, Giza, Egypt

[14]Addis Ababa University, Aklilu Lemma Institute of Pathobiology, P.O.Box 1176, Addis Ababa, Ethiopia

[15]Metabiota Inc. Cameroon.

[16]Military Veterinarian (Rtd.), P.O. Box CT2585, Accra, Ghana.

[17]Centre de Recherche en Virologie (VRV) Projet Fievres Hemoraquiques en Guinée; BP:5680 Nongo/Contéya-Commune de Ratoma, Guinea.

[18]Primate Research Center, Bogor Agricultural University, Bogor 16151, Indonesia.

[19]Faculty of Veterinary Medicine, Bogor Agricultural University, Darmaga Campus, Bogor 16680, Indonesia.

[20]Department Environment and Health, Institut Pasteur de Côte d'Ivoire, PO BOX 490 Abidjan 01, Ivory Coast.

[21]Department of Basic Medical Veterinary Sciences, College of Veterinary Medicine, Jordan University of Science and Technology, Jordan.

[22]Molecular Biology Laboratory, Institute of Primate Research, Nairobi, Kenya, Department of Biochemistry, University of Nairobi, Nairobi, Kenya.

[23]Conservation Medicine, Selangor, Malaysia.

[24]Wildlife Conservation Society (WCS) - Vietnam Program, 1302, 57 Lang Ha, Hanoi, Vietnam

[25]Center for Molecular Dynamics Nepal (CMDN), Thapathali -11, Kathmandu, Nepal.

[26]Regional Headquarters, Mountain Gorilla Veterinary Project, Musanze, Rwanda.

[27]Université Cheikh Anta Diop, BP 5005, Dakar, Sénégal.

[28]Metabiota, Inc. Sierra Leone, Freetown, Sierra Leone.

[29]Department of Veterinary Medicine and Public Health, College of Veterinary Medicine and Biomedical Sciences, Sokoine University of Agriculture, Morogoro, Tanzania.

[30]Thai Red Cross Emerging Infectious Diseases Health Science Centre, WHO Collaborating Centre for Research and Training on Viral Zoonoses, King Chulalongkorn Memorial Hospital, Faculty of Medicine, Chulalongkorn University, Bangkok, Thailand.

[31]Wildlife Conservation Society (WCS) – Bolivia Program, c. Gabino Villanueva 340, La Paz, Bolivia.

[32]Facultad de Medicina Veterinaria y Zootecnia, Universidad Nacional Autónoma de México, México City, 04510, Mexico.

[33]Centro de Biodiversidad y Genética, Universidad Mayor de San Simón, Cochabamba, Bolivia.

[34]Laboratório de Epidemiologia e Geoprocessamento (EpiGeo), Instituto de Medicina Veterinária (IMV) Universidade Federal do Pará (UFPA), BR-316 Km 31, Castanhal, PA 69746-360, Brazil.

[35]Department of Microbiology, Sanjay Gandhi Post Graduate Institute of Medical Sciences, Uttar Pradesh, India.

[36]Melbourne Veterinary School, Faculty of Veterinary and Agricultural Sciences, University of Melbourne, Werribee, VIC 3030, Australia

[37]Nyati Health Consulting, 2175 Dodds Road, Nanaimo V9X0A4, Canada.

*Correspondence to: pspandit@ucdavis.edu, ckjohnson@ucdavis.edu

**Contributed equally to the manuscript

Abstract: 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 host 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.

One Sentence Summary: Potential host range and spillover risk for novel viruses can be predicted using a network informed by known virus-host associations.

Main

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 within the context of a cluster of unusual cases, public health interventions to prevent or contain an epidemic face major challenge. 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 risk of zoonotic transmission of newly discovered animal viruses in the face of very limited data. Here we analyze human, domesticated animal, and wild animal surveillance and viral discovery data collected from 2009-2019, as part of a consortium led One Health project aimed at strengthening pandemic threat detection capabilities in Africa, Asia, and Latin America[1]. 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 established to assess potential risk, including host range or plasticity of viruses and integration of virus and ecological characteristics with expert opinion[2-5]. Using an analysis of the host-virus network we were able to quantify risk of zoonotic transmission for 531 out of 944 novel animal viruses.

Patterns observed across host-virus networks have been used to understand virus sharing among vertebrate species[3,6,7], and predict cryptic links between mammalian, and avian hosts and their viruses[8-10]. Host-virus network linkages can be informed by virus traits, virus biogeography, host ecological niches, and propensity for host sharing among viruses[10,11]. 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[2]. Further exploration of these networks can aid in estimating the host plasticity of viruses, an important characteristic associated with zoonotic potential [2,3]. Unfortunately, systematically collected surveillance data to parameterize and validate these models have been missing[4]. 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 in 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 (G_c;Fig. 1). 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 sharing of hosts[12]. 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 ($G_{predicted}$, Fig. 1) formed due to inclusion of novel viruses and their predicted linkages. We also predicted taxonomic order of the probable host 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 an edge 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.[5], Pandit et al.[4], and Johnson et al.[13] 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, 2b, 2e, 2i) for *Filoviridae*, *Flaviviridae*, *Hantaviridae*, and *Orthomyxoviridae* families were statistically different from the mean distribution (Kolmogorov-Smirnov, $p < 0.05$). Furthermore, after accounting for sampling bias for individual viruses using PubMed hits, 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 represented virus families. Viruses from the *Flaviviridae* family also had higher betweenness centrality ($p = 0.01$) indicating more connections based on shared host species (Fig 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 [3,5]. Viruses from *Nairoviridae* (p = 0.01) and *Rhabdoviridae* ($p = 0.01$) families (Fig S6) were significantly more clustered together than viruses from other families.

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)[1]. To predict associations between novel viruses nodes related to sharing 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 high performance in predicting the presence of links formed due to sharing of hosts between two virus nodes in the network. The binary model performed well in predicting sharing of viruses (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 weightage to the precision, we decided to use 0.7 as an optimum threshold for detecting a positive link between two nodes (viruses). The performance of the multilabel model varied for taxonomical orders, with higher moderate performance for predicting taxonomical orders and groups of 'humans' and Cetartiodactyla (Fig S7, Fig S9). For 531 novel viruses, we identified 184,055 possible links to new hosts (based on optimum probability threshold of 0.7 identified for the binary model) to generate the predicted network ($G_{predicted}$, Fig. 1, Fig S7a). For these predicted links, between two viruses, the multiclass model was able to to estimate 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. Empirical biological networks are rarely scale-free (network with large hubs

and showing a power-law distribution for degree)[14] but recent host-based host-virus unipartite networks have shown scale-free nature where models with power-law distributions showed the best fit for host-parasite networks[15]. 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 predicted network ($G_{predicted}$) had longer tails at network level (Kolmogorov-Smirnov, $p < 0.05$) as well as at virus family level for degree (Fig. 2a, e, f) and betweenness centrality (Fig. 2 b, i, j) distributions than the observed network (G_c). 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 S10).

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 ($p = 0.38$, Fig S11) and betweenness centrality ($p = 0.21$, Fig S12), but did significantly affect the eigenvector ($p<0.05$, Fig S13) and clustering coefficient ($p<0.05$, Fig S14) 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 higher degree ($p < 0.001$, Fig. 2C, Fig S11), betweenness ($p = 0.02$, Fig. 2D, Fig S12), and eigenvector ($p < 0.001$) centralities in the predicted network compared to newly discovered viruses in all sixteen other virus families (Fig. 2 C, D, G). In additional, the raw detection data showed significantly higher host diversity for novel coronaviruses with a mean of 2.02 (SD ± 2.03, n = 114) unique host species (maximum of 15 species) compared to 1.22 (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 breadth[16]. 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 newly detected viruses, we developed prioritization metrics based on multiclass model predicted human links for known viruses that inform on the ecological and evolutionary tendencies for spillover. 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 S15) but also were predicted to have large numbers of human links in the predicted network ($G_{predicted}$). Novel members 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 relative comparison of zoonotic risk for the newly detected viruses, a prioritization metric 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 S 16, $p < 0.001$) 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 that have numerous hosts, such as Rabies virus and West Nile virus, and consequently resulted in a prioritization score of zero. Fig. 3A-D shows the top ten and bottom five novel viruses from four virus families for relative comparison based on the prioritization metric (Fig S17-23). 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 South-east 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 new virus discoveries would only requires the data on hosts and virus families 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 [10]. Novel viruses discovered in rodents, bats, primates, and other mammalian hosts that were sampled from sites in close association with people, or at high-risk interfaces that can facilitate disease transmission in urban and rural settings[1,13]. Additional surveillance across a broader taxonomic range is essential to gain additional insight 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 for newly discovered viruses (Supplementary Data File 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 *Miniopteridae* 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 *Miniopteridae* bats. Representation of these surveillance targets through these clusters highlights host predilection of novel coronaviruses and indicates the preferential sharing of hosts by the novel coronaviruses. 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 Fig S24-S33 and supplementary data file 1.

Grange et al developed a tool that ranks viruses for animal to human spillover using a risk-based approach validated inputs by various experts from the field of virology, epidemiology and ecology[2]. Our approach, on the other hand, quantifies the risk of spillover agnostically and informs 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 predictions, 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 data[10,17] and genome sequences[18], 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. Our virus-centric approach (virus as nodes and edges as shared hosts) showed improved performance over previous host-centric models[17]. Our network approach presents some limitations specifically for viruses that have been detected in species with limited surveillance effort 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 detection of viruses in samples collected from host species) with these associations requiring further to understand relationship between viruses and hosts that might 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.

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 is included as features for prediction[11]. These tools will improve with the as well as the discovery of new viruses and further surveillance [20], 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 for the association databases included data shared by Olival et al.[5], Pandit et al.[4], and Johnson et al.[13]. 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 al[5]. In addition, a query in GenBank was run to parse out hosts reported for each GenBank submission for viruses presented in each of these three databases. Initially, for each virus name, taxonomic ID was identified using e*ntrez.esearch* function in biopython package. The taxonomic ID helped identify ICTV lineage and associated data in PubMed. 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 bias[4,13]. The virus-host association data source is presented in supplementary code and data files (https://zenodo.org/record/5899054)..

$$\begin{aligned} \boldsymbol{searchterm} = (&+\boldsymbol{virus\_name} \\ &+ [\boldsymbol{Title/Abstract}])\ \boldsymbol{AND}\ (\boldsymbol{host\ OR\ hosts\ OR\ reservoir\ OR\ reservoirs\ OR} \\ &\boldsymbol{wild\ OR\ wildlife\ OR\ domestic\ OR\ animal\ OR\ animals\ OR} \\ &\boldsymbol{mammal\ OR\ bird\ OR\ birds\ OR\ aves\ OR\ avian\ OR\ avians} \\ &\boldsymbol{OR\ vertebrate\ OR\ vertebrates\ OR\ surveillance\ OR\ sylvatic}) \end{aligned}$$

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* 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 S. 31).

Development of G_c

a. Centrality measures of observed network (G_c)

To test if centrality measures (degree centrality, betweenness centrality, eigenvector centrality, clustering coefficient) for viral nodes in the observed network (G_c) 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 networks [21,22]. For each viral family, distributions of centrality measures (degree centrality, betweenness centrality, and eigenvector centrality) and clustering coefficient within the observed network (G_c) 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.

$$centrality\ measure = \beta_0 intercept + \beta_1 Viral\ family_{categorical} + \beta_2 PubMed\ hits$$

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 ($\boldsymbol{G_c}$), multiple network topological features for all node pairs were calculated. The following are topographical network features calculated.

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. Higher Jaccard index represents similar host predilection.
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[23]. The concept of the Adamic Adar index is a weighted common neighbors for viruses in the network. Within network prediction, the index assumes that viruses with large neighborhoods have a less significant impact while predicting a connection between two viruses compared with smaller neighborhoods.

Both Jaccard and Adamic Adar coefficients have been routinely used for generalized network prediction[24].

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 coefficients: 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 k[25].
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 of 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 *v* is the fraction of nodes it is connected to. The feature that we used for training the supervised learning model was the absolute difference between degree centralities of two nodes. Unlike the difference in the 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[26]. A binary feature variable was generated to describe if both the nodes in the pair were

part of the same cluster or not. If both viruses are from the same cluster, it represents similar host predilection than when both viruses are not from the same cluster hence accounting for the evolutionary predilection of viruses (or virus families) to infect a certain type of hosts.

Pearson's correlation coefficients were calculated to identify highly correlated features and for choosing features for model training (Fig. S32). Virological features included in model training were categorical variables describing the virus family of both the nodes in the pair, followed by a binary variable if both the viruses belong to the same virus family. During the model development, PubMed hits generated three predictive features for each pair of viruses on which model training and predictions were conducted. These included two features representing PubMed hits for the two viruses in the pair ($PubMed_{V1}$, $PubMed_{V2}$) and the absolute difference between $PubMed_{V1}$ and $PubMed_{V2}$ to account for sampling bias differences between two viruses.

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 model 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 hypertuning and their performance are provided in the supplementary 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 ($\boldsymbol{G_t}$) was constructed, assuming that this represented the current observed status of the virus-host community. For all possible pairs in $\boldsymbol{G_o}$ (both that sharing and not sharing any hosts) ten topographical 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 if virus pairs share hosts (1,0 encoded output). The cross-validation was also used to determine the optimum decision threshold for determining binary classification (Fig S17) 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_c) based on species-level sharing of hosts and randomly dropping viruses to generate a training network (G_t) 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 taxonomical order (label) and were given a separate label than 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).

Missing links for novel viruses, binary and multiclass prediction: The wildlife surveillance data represented sampling of 99,379 animals (94,723 wildlife, 4,656 domesticated animals) conducted in 34 countries around the world between 2009-2019 (Table S6)[1]. Specimens were tested using conventional Rt-PCR, Quantitative PCR, Sanger sequencing, and Next Generation Sequencing protocols to detect viruses from 28 virus families or taxonomic groups (Table S7). Testing resulted in 951 novel monophyletic clusters of virus sequences (referred to as novel

viruses henceforth). Within 951 novel viruses, 944 novel viruses had vertebrate hosts that were identified with certainty based on barcoding methods and field identification. Host species identification was confirmed by cytochrome b (cytb) DNA barcoding using DNA extracted from the samples [27]. We predicted the shared host links between novel viruses and known viruses using binary and multiclass models in the following steps. Out of 944 novel viruses discovered in the last ten years, we were able to generate predictions for 531 novel viruses that were detected in species already classified as hosts within the network. The remaining 413 viruses were the first detection of any virus in that species and thus host associations could not be informed by the observed network ($\boldsymbol{G_C}$) data.

1. A new node representing the novel virus was inserted in the network of the observed network ($\boldsymbol{G_c}$). Using the list of species in which the novel virus was detected, new edges were created with known viruses that are also known to be found in those hosts. This generated a temporary network for the novel virus ($\boldsymbol{G_{temp}}$). If the novel virus was not able to generate any edges with known viruses, meaning the host in which they have been found were never found positive for any known virus, predictions were not performed.
2. Using $\boldsymbol{G_{temp}}$ feature values were calculated for the novel virus (betweenness centrality, clustering, and degree). For all possible pairs of the novel virus with known viruses that are not yet connected with each other through an edge in $\boldsymbol{G_{temp}}$ a feature dataset was generated (Jaccard coefficient$_{\text{(novel virus, known virus)}}$, the difference in betweenness centrality of the novel virus and known virus, if the novel virus and known virus were in the same cluster, the difference in degree centrality$_{\text{(novel virus, known virus)}}$, if the novel virus and known virus were from same virus family, the difference in PubMed hits$_{\text{(novel virus, known virus)}}$, PubMed hits for the novel virus, PubMed hits for the known virus). Studies and nucleotide sequences for novel viruses are expected to be published and shared on PubMed's Nucleotide database and in various peer-reviewed publications. Since, at the time of development of the model, data for all viruses was not shared in a format that would reflect on PubMed's database, we decided to use the number of times the virus was detected in the last ten years of wildlife surveillance. These detections will be reflected in PubMed's Nucleotide database eventually, hence we considered them as a proxy for search terms conducted for known viruses. Currently, evaluation of effects of this substitution of PubMed hits with the number of detections for novel viruses is not possible with limited data on novel viruses but needs to be reevaluated as more studies are published on these novel viruses.
3. Using this dataset for the novel virus, a binary presence of a link between the novel virus and known viruses was predicted using the trained binary model. The taxonomic order of the host link was predicted using the trained multiclass model.
4. For each possible link, the binary model predicted a probability of sharing link and the multiclass model predicted multivariate outcomes of taxonomic orders and associated probabilities. A threshold of 0.70 for the binary prediction model was used to classify if the link is present or not and only those links were explored for their corresponding multiclass model outputs.
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

485 importance. Firstly, the number of links where humans are predicted as shared hosts with known
486 viruses (n) and the average model-predicted probability of those links. A representation was
487 generated incorporating the probability and available model support in terms of number links to
488 reflect the likelihood and compare viruses relative to each other.

489 To test if virus family, the taxonomic order of hosts in which novel viruses were detected,
490 and the number of times the viruses were detected (equivalent to PubMed hits for known viruses)
491 influenced node (virus) level network centrality measures in the predicted network ($\boldsymbol{G_p}$) a linear
492 regression model was fitted with centrality measures.

493

494 $$centrality\ measure$$
495 $$= \beta_0 intercept + \beta_1 Viral\ family_{categorical} + \beta_2 Host\ Order_{categorical}$$
496 $$+ \beta_3 PubMed\ hits$$

497

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

503

504 **Prioritization score for novel viruses:** Generalized Linear Mixed Models were used to
505 understand the association effects of virus family, taxonomic order of the host and PubMed hits
506 on the number of predicted human links and mean probability of the predicted links. The models
507 were fit using *glmmTMB* and *glm* packages in R. For relative comparison of zoonotic risk and for
508 prioritizing novel viruses for further characterization, a prioritization metric was developed based
509 on the predicted probability of sharing the humans as hosts with known viruses ($p_{sharing\ humans}$)
510 and the number of predicted shared human links (n_{humans}) in the predicted network for the
511 given virus ($G_{predicted}$). Distributions for both $p_{sharing\ humans}$ and n_{humans} were normalized
512 and multiplied to generate a single score for a virus and for appropriate relative comparisons
513 between novel viruses. To understand the behavior of the prioritization score when predicting the
514 zoonotic risk of novel viruses, we also compared prioritization scores of known zoonotic and
515 non-zoonotic viruses using the Kolmogorov-Smirnov test.

516

517 **References:**

518 1 PREDICT Consortium. 2021. PREDICT Emerging Pandemic Threats Project. Dataset.
519 USAID Development Data Library. https://data.usaid.gov/d/tqea-hwmr..
520 2 Grange, Z. L. *et al.* Ranking the risk of animal-to-human spillover for newly discovered
521 viruses. *Proceedings of the National Academy of Sciences* **118**, e2002324118,
522 doi:10.1073/pnas.2002324118 (2021).
523 3 Kreuder Johnson, C. *et al.* Spillover and pandemic properties of zoonotic viruses with
524 high host plasticity. *Sci Rep* **5**, 14830, doi:10.1038/srep14830 (2015).
525 4 Pandit, P. S. *et al.* Predicting wildlife reservoirs and global vulnerability to zoonotic
526 Flaviviruses. *Nat Commun* **9**, 5425, doi:10.1038/s41467-018-07896-2 (2018).
527 5 Olival, K. J. *et al.* Host and viral traits predict zoonotic spillover from mammals. *Nature*
528 **546**, 646-650, doi:10.1038/nature22975 (2017).

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 U S A* **110**, 7738-7741, doi:10.1073/pnas.1220716110 (2013).

7 Albery, G. F. *et al.* The science of the host–virus network. *Nature Microbiology* **6**, 1483-1492 (2021).

8 Walker, J. G., Plein, M., Morgan, E. R. & Vesk, 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**, doi: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, doi:10.1371/journal.pcbi.1005557 (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**, 2260, doi:10.1038/s41467-020-16153-4 (2020).

11 Wardeh, 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. *Nature Communications* **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. *Proc Biol Sci* **287**, 20192736, doi:10.1098/rspb.2019.2736 (2020).

14 Broido, A. D. & Clauset, A. Scale-free networks are rare. *Nat Commun* **10**, 1017, doi:10.1038/s41467-019-08746-5 (2019).

15 Carlson, C. J., Zipfel, C. M., Garnier, R. & Bansal, S. Global estimates of mammalian viral diversity accounting for host sharing. *Nat Ecol Evol* **3**, 1070-1075, doi:10.1038/s41559-019-0910-6 (2019).

16 Banerjee, A., Mossman, K. & Baker, M. L. Zooanthroponotic 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.* Optimizing predictive models to prioritize viral discovery in zoonotic reservoirs. *bioRxiv*, 2020.2005. 2022.111344 (2021).

18 Mollentze, N., Babayan, S. & Streicker, D. Identifying and prioritizing potential human-infecting viruses from their genome sequences. *bioRxiv*, 2020.2011. 2012.379917 (2021).

19 Mollentze, N. & Streicker, D. G. 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).

20 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, doi:10.1098/rstb.2011.0354 (2012).

21 Kossinets, G. & Watts, D. J. Empirical analysis of an evolving social network. *Science* **311**, 88-90, doi:10.1126/science.1116869 (2006).

22 Muchnik, L. *et al.* Origins of power-law degree distribution in the heterogeneity of human activity in social networks. *Sci Rep* **3**, 1783, doi:10.1038/srep01783 (2013).

23 Adamic, L. A. & Adar, E. Friends and neighbors on the web. *Social networks* **25**, 211-230 (2003).

24 Lü, L. & Zhou, T. Link prediction in complex networks: A survey. *Physica A: statistical mechanics and its applications* **390**, 1150-1170 (2011).

25 Barabasi, A. L. & Albert, R. Emergence of scaling in random networks. *Science* **286**, 509-512, doi:10.1126/science.286.5439.509 (1999).
26 Blondel, V. D., Guillaume, J.-L., Lambiotte, R. & Lefebvre, E. Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment* **2008**, P10008 (2008).
27 Irwin, D. M., Kocher, T. D. & Wilson, A. C. Evolution of the cytochrome b gene of mammals. *J Mol Evol* **32**, 128-144, doi:10.1007/BF02515385 (1991).

Acknowledgments:

Funding: This work was supported by the United States Agency for International Development (USAID) Emerging Pandemic Threat PREDICT program (Cooperative Agreement nos. GHN-A-00-09-00010-00 and AID-OAA-A-14-00102). . P.S.P. is also supported by R supported by the National Institute Of Allergy And Infectious Diseases of the National Institutes of Health under Award Number U01AI151814. The content is solely the responsibility of the authors and does not necessarily represent the official views of the 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, South Sudan, Tanzania, Thailand, Uganda, and Vietnam for permission to conduct this study, and the field teams and collaborating laboratories that performed sample collection and testing.

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., C.K.J., S.J.A, T.G, K.L.O, J.A.K.M, M.M.D., N. R. G., B. B., W. 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. A. C., A. L., C. L., T. O'R., S. O., L. K., P. M., A. P., C. D. de P., D. Z., M. V., M. LB., D. MI., A. I., V. D., M. M., Z. S., P. M., M. A., N. K., U. T., S. B.N., A. C., J. P., K. C., E. A.B., J. K., S. S., J. D., T. H., E. S., O. A., D. K., J. N., D. N., A. G., Z. S., S. W., E. A. R., B. S., G. S., L. F. A., M. R. S., T. N. D., P. L. H., D. O. J., K. S., A. F., S. M., W. K., P. D., J., and PREDICT Consortium collected data, wrote and revised the manuscript.

Competing interests: Author declare no competing interests.

Data and materials availability: Data and code reported in this paper are available at https://zenodo.org/record/5899054 and https://data.usaid.gov/Global-Health-Security-in-Development-GHSD-/PREDICT-Emerging-Pandemic-Threats-Project/tqea-hwmr

Figures and Tables

Fig. 1. Modeling workflow: The figure shows modeling procedure and methods implemented in the study. Orange dot represents a known virus in the observed (G_c) and predicted networks ($G_{predicted}$), blue dots represent novel viruses in the predicted network ($G_{predicted}$). Virus-host networks: G_c, represents a unipartite observed network of known zoonotic and non-zoonotic viruses with nodes representing viruses and edges representing shared hosts. G_$G_{predicted}$represents the predicted

unipartite network generated after predicting possible linkages between 531 novel viruses (white) and known viruses. The node size is proportional to the betweenness centrality.

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 (C) degree, (D) betweenness centrality, (G) eigenvector centrality, and (H) clustering based on the predicted network formed by the binary prediction model.

Fig. 3: Prioritization metrics for novel viruses to understand zoonotic risk: Top ten and bottom five newly discovered viruses from six virus families (A-F) with the virus prioritization scores based on multiclass model predictions. Annotations show the score and support represented by number of human links predicted.

Fig. 4: Surveillance targets for novel coronaviruses based on predicted sharing of hosts with known viruses. Red color represents the evidence towards species in the taxonomic family (cumulative probability) with darker red color indicating higher number of species occurrences from taxonomical families adjusted by model predicted probability. A) shows clustering of PREDICT coronaviruses by host, and B) focuses on coronaviruses found in bats. Clustering is based on the Bray-Curtis dissimilarity index**.**

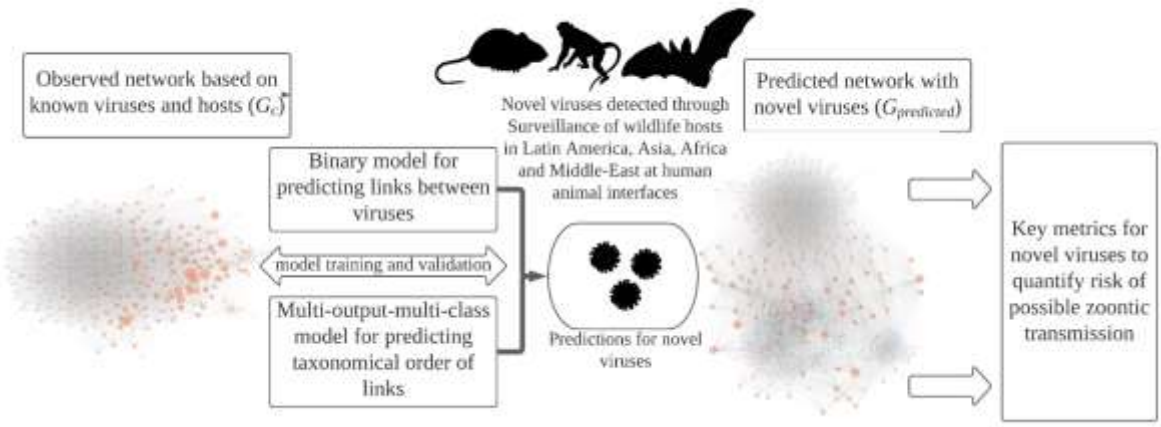

Fig. 1: Model prediction workflow: The figure shows modeling procedure and methods implemented in the study. Orange dot represents a known virus in the observed (G_c) and predicted networks ($G_{predicted}$), blue dots represent novel viruses in the predicted network ($G_{predicted}$). Virus-host networks: G_c, represents a unipartite observed network of known zoonotic and non-zoonotic viruses with nodes representing viruses and edges representing shared hosts. G_$G_{predicted}$ represents the predicted unipartite network generated after predicting possible linkages between 531 novel viruses (white) and known viruses. The node size is proportional to the betweenness centrality.

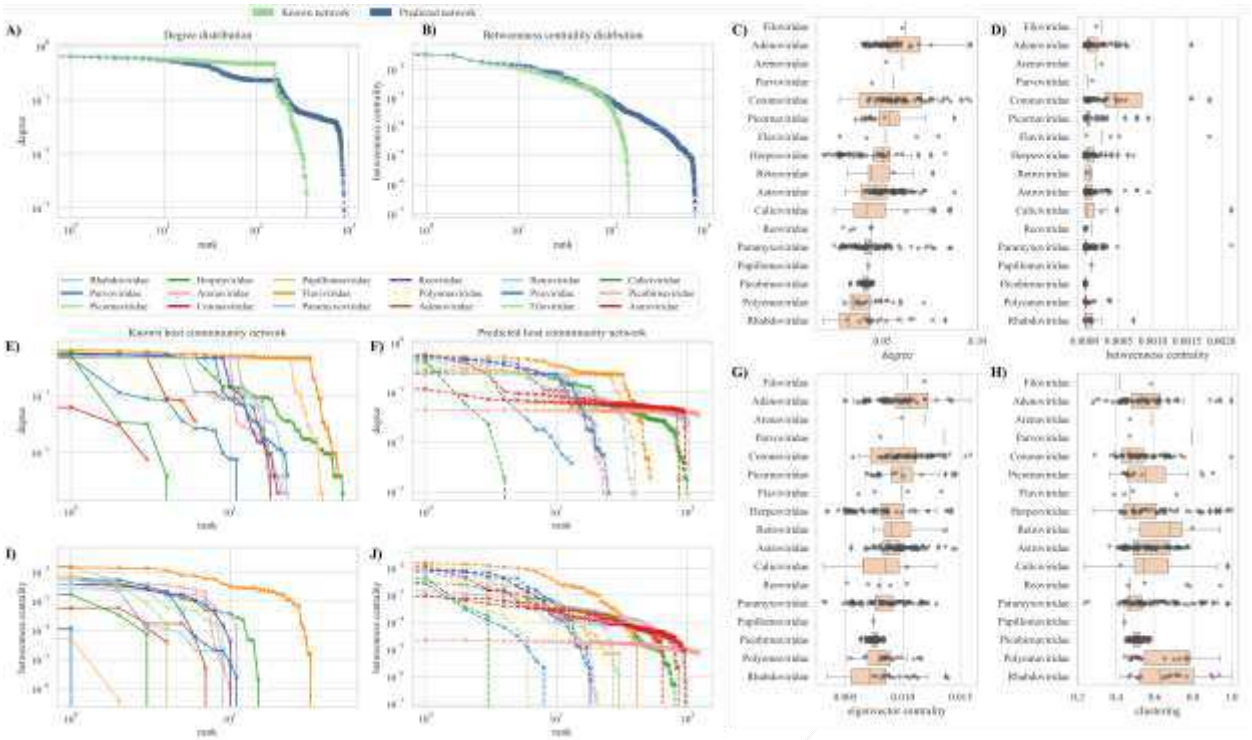

*Fig. 2**: Predicting missing links between virus-host communities.** Distribution shapes of degree (A) and betweenness centrality (B) for the observed and predicted network. Degree distributions for virus families in observed and predicted networks are shown in (E) and (F). Similarly, shapes of betweenness centrality for virus families in observed and predicted networks are shown in (I) and (J). Right panels show boxplots for novel virus families describing (C) degree, (D) betweenness centrality, (G) eigenvector centrality, and (H) clustering based on the predicted network formed by the binary prediction model.*

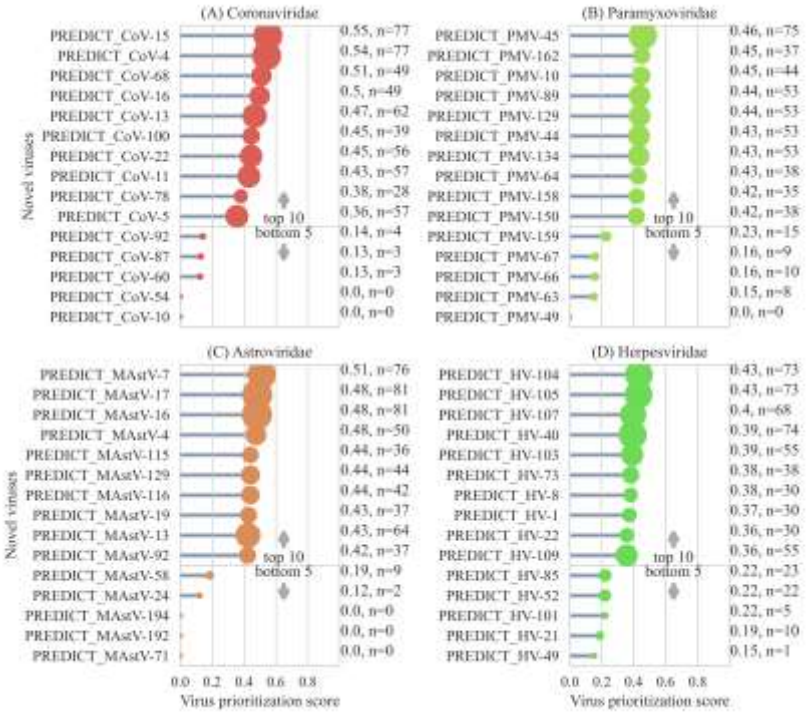

Fig. 3: ***Prioritization metrics for novel viruses to understand zoonotic risk****: Top ten and bottom five newly discovered viruses from six virus families (A-D) with the virus prioritization scores based on multiclass model predictions. Annotations show the score and support represented by number of human links predicted.*

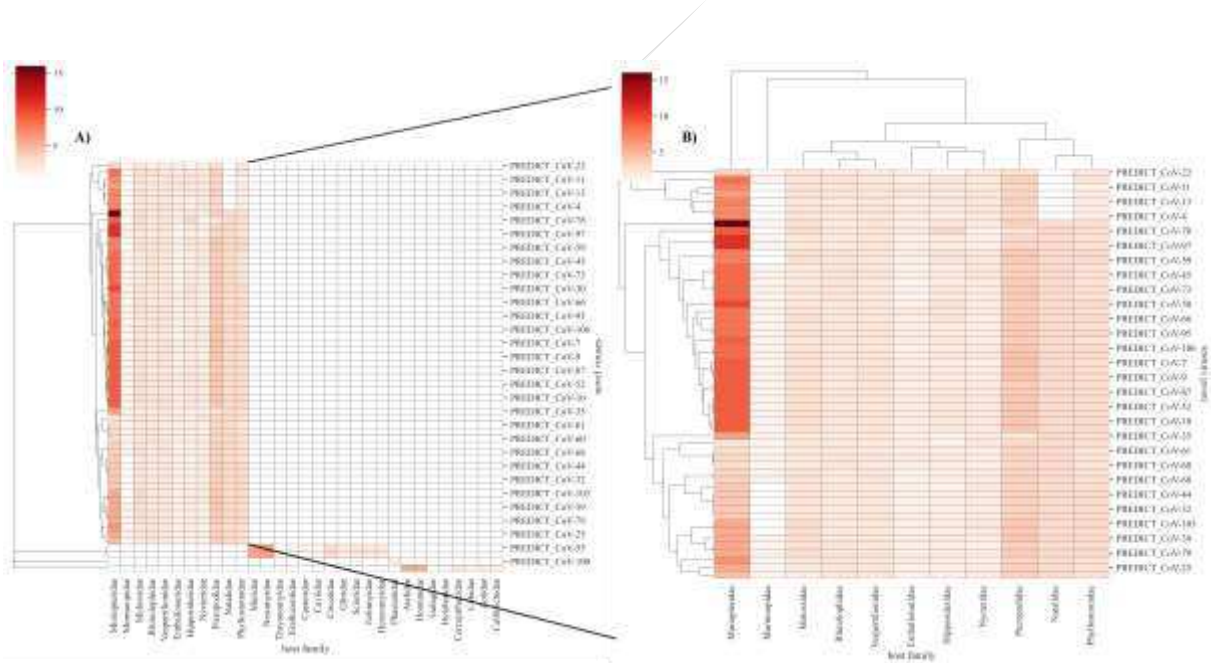

Fig. 4: Surveillance targets for novel coronaviruses based on predicted sharing of hosts with known viruses. Red color represents the evidence towards species in the taxonomic family (cumulative probability) with darker red color indicating higher number of species occurrences from taxonomical families adjusted by model predicted probability. A) shows clustering of newly discovered coronaviruses by host, and B) focuses on coronaviruses found in bats. Clustering is based on the Bray-Curtis dissimilarity index.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryMaterialsPanditJohnsonetalr1.docx
- Supplementraydatafile1.xlsx