PNAS RESEARCH ARTICLE | ECOLOGY

Wildlife susceptibility to infectious diseases at global scales

Ángel L. Robles-Fernández[a,b,1], Diego Santiago-Alarcon[c,1], and Andrés Lira-Noriega[d,1]

Edited by Nils Stenseth, Universitetet i Oslo, Oslo, Norway; received December 17, 2021; accepted July 11, 2022

Disease transmission prediction across wildlife is crucial for risk assessment of emerging infectious diseases. Susceptibility of host species to pathogens is influenced by the geographic, environmental, and phylogenetic context of the specific system under study. We used machine learning to analyze how such variables influence pathogen incidence for multihost pathogen assemblages, including one of direct transmission (coronaviruses and bats) and two vector-borne systems (West Nile Virus [WNV] and birds, and malaria and birds). Here we show that this methodology is able to provide reliable global spatial susceptibility predictions for the studied host–pathogen systems, even when using a small amount of incidence information (i.e., <20% of information in a database). We found that avian malaria was mostly affected by environmental factors and by an interaction between phylogeny and geography, and WNV susceptibility was mostly influenced by phylogeny and by the interaction between geographic and environmental distances, whereas coronavirus susceptibility was mostly affected by geography. This approach will help to direct surveillance and field efforts providing cost-effective decisions on where to invest limited resources.

disease risk | emerging infectious diseases | global epidemiology | machine learning | One Health

A lasting challenge in disease ecology is to determine pathogen emergence risks (1, 2). This task has become cumbersome because human impacts may uncouple biogeographical patterns, altering the ecological and evolutionary dynamics of host–pathogen systems (3, 4). The rate of human impacts on natural environments has increased steadily during the last 2 centuries, opening opportunities for novel host–parasite associations via host-switching (5, 6). The number of emerging infectious diseases in both humans and nonhuman organisms has increased during the last 3 decades (7–9), particularly across the human–domestic–wildlife interface, for instance, via biological invasions (10, 11). In the case of human diseases, emergence events are more likely to take place in tropical regions with warmer and humid climates, in areas with higher host diversity—particularly mammals such as rodents, bats, and nonhuman primates—and in regions where there is a higher land use change rate toward agroecosystems and urbanization (6, 12–15).

Even when viruses and bacteria are the most common zoonoses (7, 8), it is difficult to predict what kind of pathogen will produce the next medical and/or veterinary challenge (8, 16). Thus, it would be ideal to provide a general framework applicable to a diverse array of host–pathogen systems that considers complete host assemblages, together with their geographic, environmental, and phylogenetic contexts in order to predict host risks and to discover potential novel reservoirs (17). The most recent emergence of COVID-19 (i.e., severe acute respiratory syndrome coronavirus 2 [SARS-CoV-2]) human pandemic (18) is a clear example of a global need to understand how zoonotic pathogens distribute among related taxa that are known to transmit them (e.g., bat and rodent host species), as well as to determine how likely it is for other related host species to share zoonosis based on geographic, environmental, and/or phylogenetic distances. Studies spanning a diverse array of host–parasite interactions have demonstrated that a large amount of variation in host breadth is explained by host phylogenetic relationships (19–24), environmental variables (13, 23, 25), and the richness of both parasites and host clades (15, 26). Although studies focusing on viruses and mammals have provided spatial information on emergent diseases hot spots and their environmental correlates (12, 13, 27), their results are not necessarily generalizable to other host–pathogen systems. Here we contrast predictions of multihost–multipathogen assemblages estimated from geographic, environmental, and phylogenetic distances via machine learning statistical protocols in order to forecast infection susceptibility to potential hazards (1). We implemented our approach to two different Diptera-borne parasite systems (i.e., avian malaria and West Nile Virus [WNV] infecting birds) and to directly transmitted coronaviruses infecting bats, in order to determine the generality and applicability of our procedure. The proposed predictive framework (Fig. 1), in addition to known hosts, also helps to identify potential host species where the pathogen has not been previously detected, or it has not been considered in field screening projects, or it is not necessarily the host species with the highest incidence.

Significance

Previous studies have investigated how environmental, phylogenetic, and geographic variables determine pathogen infection, particularly for human zoonosis. Yet, none of those previous studies have provided a methodology that can be applied to a broad array of host–pathogen systems. Here, we provide a machine learning approach that can integrate different explanatory variables and be applied to any multihost–multipathogen system. Our results agree with the known ecology of each analyzed system and provide a tool that can help discovering potential host species and novel geographical hot spots for a pathogen. Thus, it can help guiding sampling decisions in terms of both host species and geographical locations. Finally, this tool can be applied at different spatial scales with few incidence data.

Author affiliations: [a]Facultad de Física, Universidad Veracruzana, 91000 Xalapa, México; [b]School of Life Sciences, Arizona State University, Tempe, AZ 85281; [c]Department of Integrative Biology, University of South Florida, Tampa, FL 33620; and [d]Red de Estudios Moleculares Avanzados, Instituto de Ecología, A.C., 91073 Xalapa, México

Author contributions: Á.L.R.-F., D.S.-A., and A.L.-N. designed research; Á.L.R.-F., D.S.-A., and A.L.-N. performed research; Á.L.R.-F. contributed new reagents/analytic tools; Á.L.R.-F., D.S.-A., and A.L.-N. analyzed data; and Á.L.R.-F., D.S.-A., and A.L.-N. wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission.

[1]To whom correspondence may be addressed. Email: a.l.robles.fernandez@gmail.com, santiagoalarcon@usf.edu, or aliranoriega@gmail.com.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2122851119/-/DCSupplemental.

Published August 22, 2022.

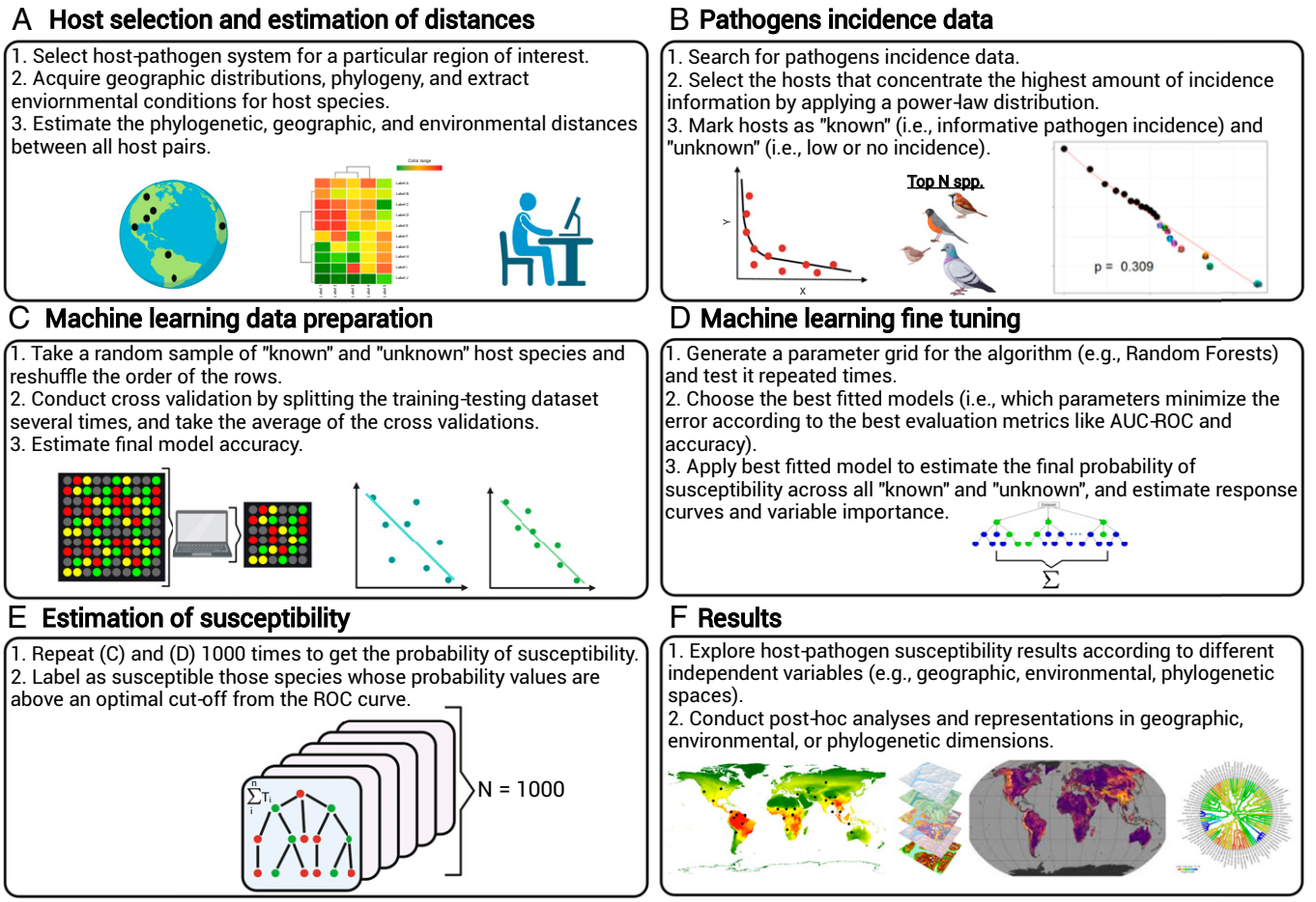

Fig. 1. Workflow of steps to estimate the susceptibility of wildlife to diseases. The steps within each of the six main steps (*A*–*F*) summarize the proposed methodology to estimate wildlife susceptibility to infectious diseases based on machine learning.

This endeavor is possible due to the implementation of classification algorithms via machine learning protocols that aid in identifying the influence of independent variables on known cases of pathogen incidence. This allows us to estimate a probability of susceptibility on potential host species as a function of their similarity in the combination of independent variables. This framework has several advantages compared to more traditional approaches: it is statistically robust even for biased pathogen incidence data, the outcome is consistent with previous biological knowledge of the host–pathogen system under study, and by considering host similarity under multiple dimensions simultaneously (e.g., geographic, environmental, and phylogenetic) we are able to infer probabilistically the susceptibility of a host species to a pathogen.

Avian malaria is a vector-borne disease generated by parasites of the genus *Plasmodium* (Haemosporida: Plasmodidae) (28), in particular by the species *Plasmodium relictum* that is a host generalist global invasive pathogen—responsible, along with habitat destruction and transformation, for the extinction of many Hawaiian endemic birds (29)—composed of five genetic variants (30). Avian hemosporidians have faced few geographical barriers over their evolutionary history, readily dispersing across biogeographical regions (31). Furthermore, hemosporidians are able to infect a large array of host species, but they usually infect phylogenetically closely related hosts and not necessarily hosts with similar ecological niches (21). Temperature and rainfall seasonality predict a higher parasite host specialization and assemblage uniqueness (32); for *Plasmodium*, there is a negative association between maximum temperature and phylobeta diversity, suggesting that as temperature increases, communities become more homogeneous (33). Thus, for avian malaria we predict that phylogenetic and environmental variables must be the most important factors predicting bird species susceptibility. WNV (Flaviviridae) is also a vector-borne pathogen transmitted mostly by mosquitoes of the genus *Culex*; it is composed of different genetic strains with varying degrees of virulence (34). WNV has well-established sylvatic cycles involving birds as primary host species and reservoirs—particularly Passeriformes, Charadriiformes, Falconiformes, Accipitriformes, and Piciformes (35, 36). Although humans and other nonhuman mammals are regarded as incidental or dead-end hosts (36), WNV has been detected infecting a large array of mammal orders (e.g., Chiroptera, Carnivora, and Artiodactyla), with a tendency to specialize in rodents (34, 37). Similar to avian malaria, WNV is geographically widespread—aided via migratory birds (38)—and it is rather limited by climatic conditions (e.g., temperature and precipitation) and land use type that affect vector survival and reproduction (39, 40). Thus, for WNV we predict that environmental variables first, followed by phylogenetic host relationships, would be the most important factors predicting susceptibility of bird species. Coronaviruses are a highly diverse group of directly transmitted parasites, where zoonotic pathogens creating severe human health effects (e.g., SARS-CoV-2 and Middle East respiratory syndrome) are but a small proportion of the diversity recorded in bats (41–43). This is actually a general trend for zoonoses that have affected human populations during the last 3 decades (i.e., outbreaks are generated by a small fraction of all the zoonotic richness recorded in humans; about 80% of cases from

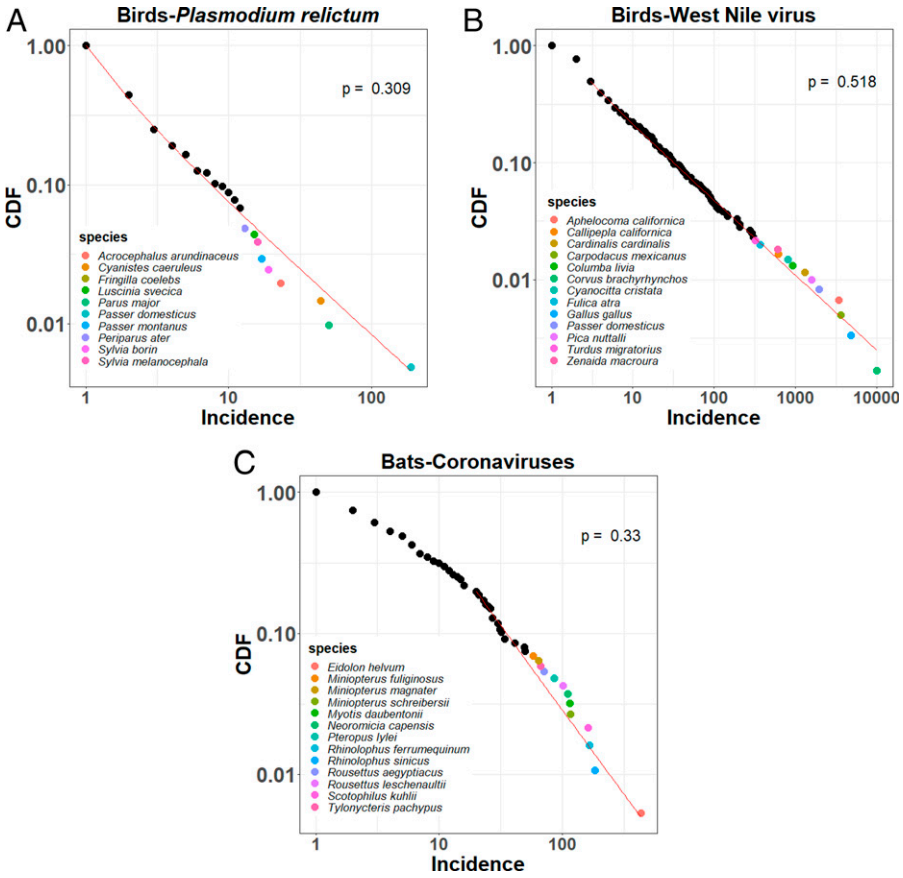

Fig. 2. Complementary cumulative density function (CDF) of host–pathogen incidence for the three studied systems: (*A*) Birds-*Plasmodium relictum*, (*B*) Birds-West Nile Virus, and (*C*) Bats-coronaviruses. The red line in the plots shows the theoretical incidence that corresponds to the power law distribution, and the colored dots correspond to the species with highest amount of incidence information useful in model calibration. *P* value within the plots indicates that none of the complementary CDFs of host–pathogen systems statistically departed from a power law distribution.

1980 to 2010 are generated by 20% of recorded zoonoses) (8). There is a positive correlation between coronavirus diversity and bat diversity, and there is also coronaviruses assemblage turnover determined by beta phylodiversity of bats across biogeographical regions (41). Furthermore, coronaviruses readily switch hosts creating a challenge to discover the intermediate hosts acting as reservoirs and aiding in transmission (44); this is particularly relevant when the system includes migratory bat species that are able to connect wild communities across large distances (27). The bat coronaviruses system does not seem to be affected by climatic conditions; instead, its distribution and host breadth are determined by ecological (e.g., gregariousness, diet, and dispersal/migration) and phylogenetic factors, where its expansion may be aided by sympatric rodent species (27, 45). Therefore, we expect that geographical and phylogenetic variables are the most important in determining coronavirus host susceptibility. In this study, we showed that our methodology is able to provide reliable spatial infection risk predictions by using a small amount of information from a host–parasite assemblage (i.e., <20% of host–pathogen incidence information). We attempt to provide a generalizable methodology across host–parasite systems identifying host community hot spots of infection risk at a global scale.

Results

Host–Pathogen Incidence. We found that the cumulative distribution function of pathogen incidence in all three assemblages follows a power law distribution despite being different epizootiological systems (Fig. 2). This allowed us to select the largest amount of incidences for model calibration by using the lowest amount of data (see *Host–Pathogen Data* subsection in *Materials and Methods*), while still preserving the generality and explanatory power of the independent variables despite biases in sampling.

Accuracy of the Models. In Fig. 3 we show the accuracy and the area under receiver operating characteristic curve (AUC–ROC) for each of 1,000 runs of random forest models. Both metrics show an acceptable performance comparing the median of the 1,000 runs in the three host–pathogen systems.

Species Susceptibility and Variable Importance. We found that for the three systems, host species with the highest pathogen incidence were not necessarily those predicted to be the species with the highest probability of being susceptible (Tables 1–3 show top 10 species by incidence and susceptibility; see https://doi.org/10.5281/zenodo.6510454 for all species results). We discovered susceptible host species that were not part of model construction or calibration; furthermore, some host species with low incidence values were predicted to have a high degree of susceptibility. Thus, currently known incidence values are not necessarily determinant of host susceptibility to pathogens in a given system. Yet, each host–pathogen system differs in the variables that are important to determine susceptibility (i.e., geography, environment, and phylogeny). Avian malaria was mostly affected by environmental distance between hosts, followed by a combination of host geographic distribution and host phylogenetic relationships; WNV was influenced by a combination of the three sets of predictors, while bat coronaviruses were mostly affected by the geographic distribution of susceptible host species (Fig. 4). From a

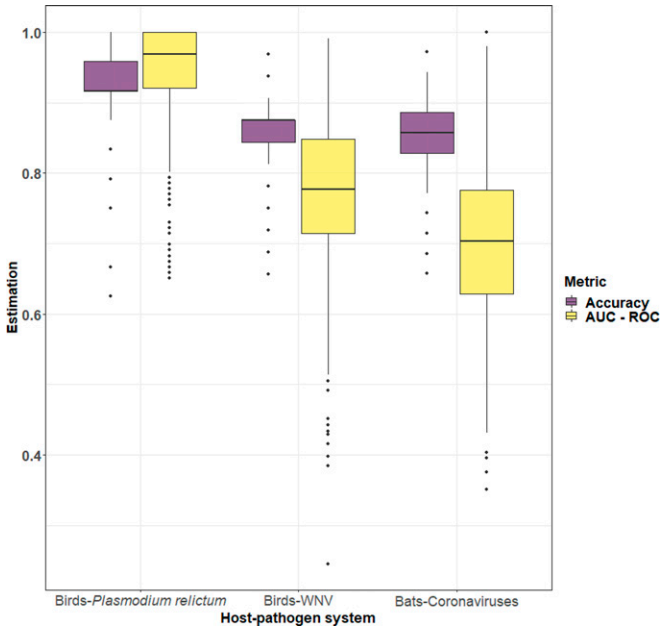

Fig. 3. Accuracy and area under receiver operating characteristic curve (AUC–ROC). Each model was run on a balanced resample of infected known hosts and potential unknown hosts. We select the best model after the cross-validation in the training dataset and perform the metrics on the testing dataset for each sample. Accuracy (purple) is the proportion of well-classified points in the test dataset after the cross-validation. AUC of the ROC curve (yellow) is the proportion of the true positive rate against the false positive rate, where values between [0.75, 0.9) are considered as optimal performance.

biogeographic perspective, avian malaria hosts showed more susceptibility across the Palearctic region, followed by the Afrotropical and Indo-Malay regions; the WNV hosts had higher susceptibility in the Palearctic and North American regions and lower toward the Indo-Malay region; finally, the bat coronavirus hosts' susceptibility was strongest in the Indo-Malay and the Afrotropical regions and the southwestern portion of the Palearctic region.

We found that the point patterns of the empirical data followed the richness of susceptible species (Fig. 4). The point pattern analyses were conducted using the georeferenced information from the databases and are considered independent to the model given that these points never formed part of the implementation of the model. The spatial point density fits better in the case of avian malaria and WNV, whereas for the coronaviruses system the uncertainty is due both to the use of a higher taxonomic level (i.e., family) and to the fact that the geographical data for the empirical incidence information were obtained using geopolitical units (e.g., capital cities or regions of the study) given the lack of exact geospatial information.

Environmental Envelopes. Regarding superposition in environmental space, we present the environmental envelopes as ellipses of the six species with the highest incidence and susceptibility for the three host–pathogen systems (Fig. 5). Overall, the three systems showed a large amount of overlap in the host species with highest incidence; however, this trend was larger for the bird–malaria system in comparison to the host species of the other two systems.

In the case of environmental envelopes for susceptibility, in the bird–malaria system, species occupied a broad spectrum of the first principal component similar to the highest incidence host species, while for bird–WNV and for bat coronaviruses the most susceptible species occupied a broader spectrum of environmental space. However, this is not necessarily reflected in the global variable importance pattern (Fig. 4).

Susceptibility as a Phylogenetic Trait. In the case of avian malaria, we observed that all susceptible bird species belonged to the order Passeriformes, with particular high susceptibility in the families Fringillidae, Motacillidae, Emberizidae, and Acrocephalidae (Table 1 and Fig. 6). For the WNV, we observed that both passerine and nonpasserine birds were susceptible (Table 2 and Fig. 6). Among the passerines, the bird families with high WNV susceptibility included Turdidae, Sittidae, and Corvidae. In the case of nonpasserines, the most susceptible families included Rheidae, Anatidae, Phasianidae, Pteroclidae, Columbidae, Rallidae, and Charadriidae, among others. Finally, for the bat coronaviruses it is clear that this virus family is widespread across the Chiroptera. Bat genera with high susceptibility included *Rhinolophus*, *Eidolon*, *Rousettus*, *Pipistrellus*, *Vespertilio*, *Tylonycteris*, and *Scotophilus*, among others (Table 3 and Fig. 6).

Discussion

Pathogen species responses to ecosystem factors are not general (12, 46–51). Idiosyncratic outcomes (49, 52–55) have called for the urgent need of tools in disease ecology to forecast when and where pathogen outbreaks are likely. Here we have provided a methodological framework that considers geographical, environmental, and phylogenetic variables of host species applicable to any host–pathogen system (e.g., with direct or indirect transmission). Although we have used the procedure at global scales, the approach is applicable at different spatial scales from the landscape to the global (i.e., >10 km; sensu ref. 56), which is important because factors may not behave in the same manner as the temporal or spatial scales of analysis change (e.g., refs. 51, 55, 57). In this way, it is possible to identify those factors most relevant to each system in order to determine hosts with a higher risk of infection, including those host species that have not been sampled in field surveys or that have not been recorded as infected by the parasite under study, providing a way to manage, prevent, and mitigate pathogen risk.

As expected from knowledge on parasite life cycles, our method identified relevant factors for each system. Avian malaria—*P. relictum*—is a widely distributed and host generalist invasive species, infecting 300 bird species worldwide (30). This parasite species is composed by five genetic lineages (30). Although they are widespread across continents, there is a geographical division in their nuclear genetic variation: SGS1 and GRW11 genetic variants are mostly restricted to Europe and Asia, whereas GRW04 is the only lineage in the Americas with nuclear genetic variants that are not present in Africa, Europe, and Asia (58). Accordingly, our models correctly identified the geographical and environmental factors as the most relevant; there was also an interaction between geography and phylogeny, where *P. relictum* mostly infects passerines. Although avian malaria is a widespread generalist pathogen, previous studies have demonstrated that avian hemosporidians, even those with large host breadths, infect mostly closely related (i.e., family-level and lower taxonomic ranks) host species (e.g., refs. 21, 59). Moreover, *P. relictum* has a worldwide genetic structure that corresponds largely with continental masses, which have different avian compositions corresponding to different evolutionary histories of biogeographical regions (58, 60). Avian malaria has basically colonized all types of environments across the world, so environmental conditions can affect its prevalence only in a seasonal fashion in temperate and cold climates, where Culicidae vectors are not active during the cold months or at high-elevation environments (e.g., refs. 61, 62), which agrees with the importance of temperature seasonality and mean temperature

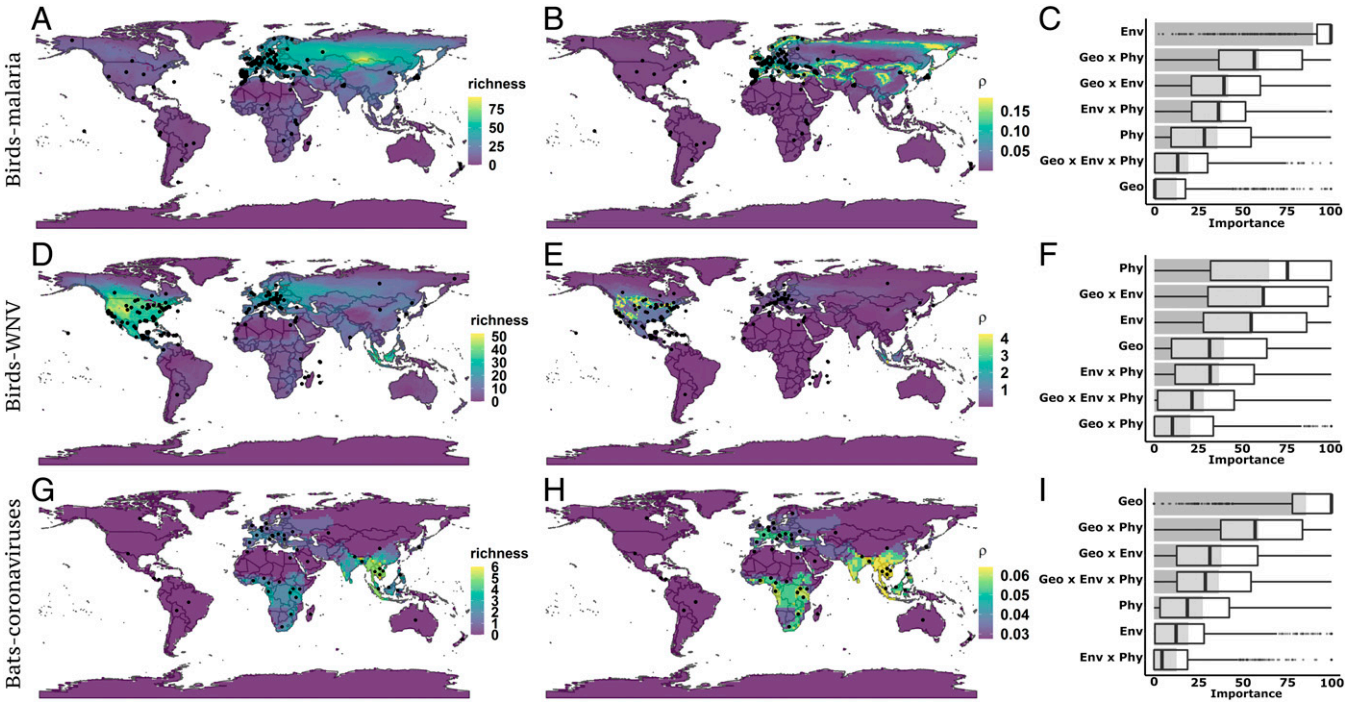

Fig. 4. Susceptibility richness maps, point intensity pattern as a function of susceptibility richness (ρ), and variable importance in each host–pathogen system. The black dots are empirical observations for the focal pathogen (*Matherials and Methods*). (*A*) Bird species richness with susceptibility value to *P. relictum*. (*B*) Point intensity pattern as a function of species susceptibility to *P. relictum* in bird species. (*C*) Summary of variable importance for bird species susceptibility to *P. relictum*. (*D*) Bird species richness with susceptibility to WNV. (*E*) Point intensity pattern as a function of species susceptibility to WNV in bird species. (*F*) Summary of variable importance for bird species susceptibility to WNV. (*G*) Bat species richness with susceptibility to coronaviruses. (*H*) Point intensity pattern as a function of species susceptibility to coronaviruses in bat species. (*I*) Summary of variable importance for bat species susceptibility to coronaviruses. In all three cases, susceptible host species correspond to susceptibility value $\geq$0.5, and the importance of the variables was determined after 1,000 random forest runs (boxplots of variable importance summarize the second, the median, and the third quartile, and the gray bar corresponds to the mean of variable importance).

on hemosporidian prevalence (32, 62). Our model corroborates the house sparrow (*Passer domesticus*) as an important species in terms of infections by *P. relictum*, which is also a widespread invasive urban bird known to outcompete resident species in invaded ranges (63) and serves as reservoir of avian malaria that can be transmitted to native birds (64, 65). Interestingly, the model did not include any of the highly infected (i.e., high-incidence) *Sylvia* spp. within the top 10 most susceptible bird species to avian malaria; instead, other species were included in the top 10 (Table 1; https://doi.org/10.5281/zenodo.6510454). Our model provides a list of highly susceptible hosts that may either have no recorded avian malaria infections (e.g., genus *Motacilla*, *Anthus godlewskii*, and *Fringilla montifringilla*) or present either high (*P. domesticus*) or low incidence (e.g., *Hirundo rustica* and *Acrocephalus schoenobaenus*; Table 1; https://doi.org/10.5281/zenodo.6510454), hence providing a guide in terms of potentially important hosts and reservoirs of avian malaria for future sampling. Regarding the geographic context, Eurasia is the region that our model predicts with the highest species richness of susceptible avian hosts, more specifically, the region encompassing southern Russia, eastern Kazakhstan, and northwest Mongolia that contains a high species richness of susceptible hosts but is clearly undersampled.

WNV is a pathogen composed of different genetic variants characterized by different degrees of virulence and considered to be of concern to birds' health, with occasional human and nonhuman mammal cases (36). In recent years, it has become clear that WNV readily infects mammals from different orders, specializing to some degree in rodents, but mammal competence to WNV is not well determined (22). This information suggests that WNV may not be limited by host phylogenetic associations, with potential to spillover across birds and mammals (e.g., ref. 66). Moreover, it should not be limited by geography given the long list of bird and mammal hosts recorded so far, many of which have broad geographical ranges—including long-distance migrants (22, 38). Although our WNV model only included bird species, it correctly identified environmental variables as a relevant factor determining host susceptibility, interacting with geography, and having a similar influence by host phylogenetic associations. WNV is an endemic pathogen of Africa that has dispersed across the world mainly in temperate Europe, North America, and Asia. Our model identified that the highest richness of susceptible bird hosts is located in North America and Eurasia. The lack of sampling is one of the reasons why WNV may have not been commonly recorded in birds of tropical regions, where urban and nonurban cycles have been established via wild and domestic birds (e.g., ref. 67). A second reason may be a high diversity of both birds and mammals in invaded tropical regions that may act as a dilution factor, particularly by infectious mosquito bites intercepted by rodents and bats that seem to be refractive to WNV infections (e.g., ref. 34). Third, WNV hosts have a different evolutionary history in invaded regions; lower phylogenetic similarity may also imply more divergent immune systems (e.g., ref. 68), reducing WNV infection success. Moreover, WNV may be competing both with closely related flaviviruses (e.g., Saint Louis encephalitis virus) and with other WNV genetic variants recently evolved in invaded areas [e.g., Florida (69)] that are better adapted to native hosts and thus are less virulent. Fourth, high temperatures interrupt the WNV transmission cycle in mosquitoes—similar to what happens with avian malaria (70), where the optimal temperature for transmission is between 24 and 25 °C (71). As global warming increases, then WNV is expected to expand toward currently cooler areas (71), something that researchers can start exploring by using real-time tools like the one applied to

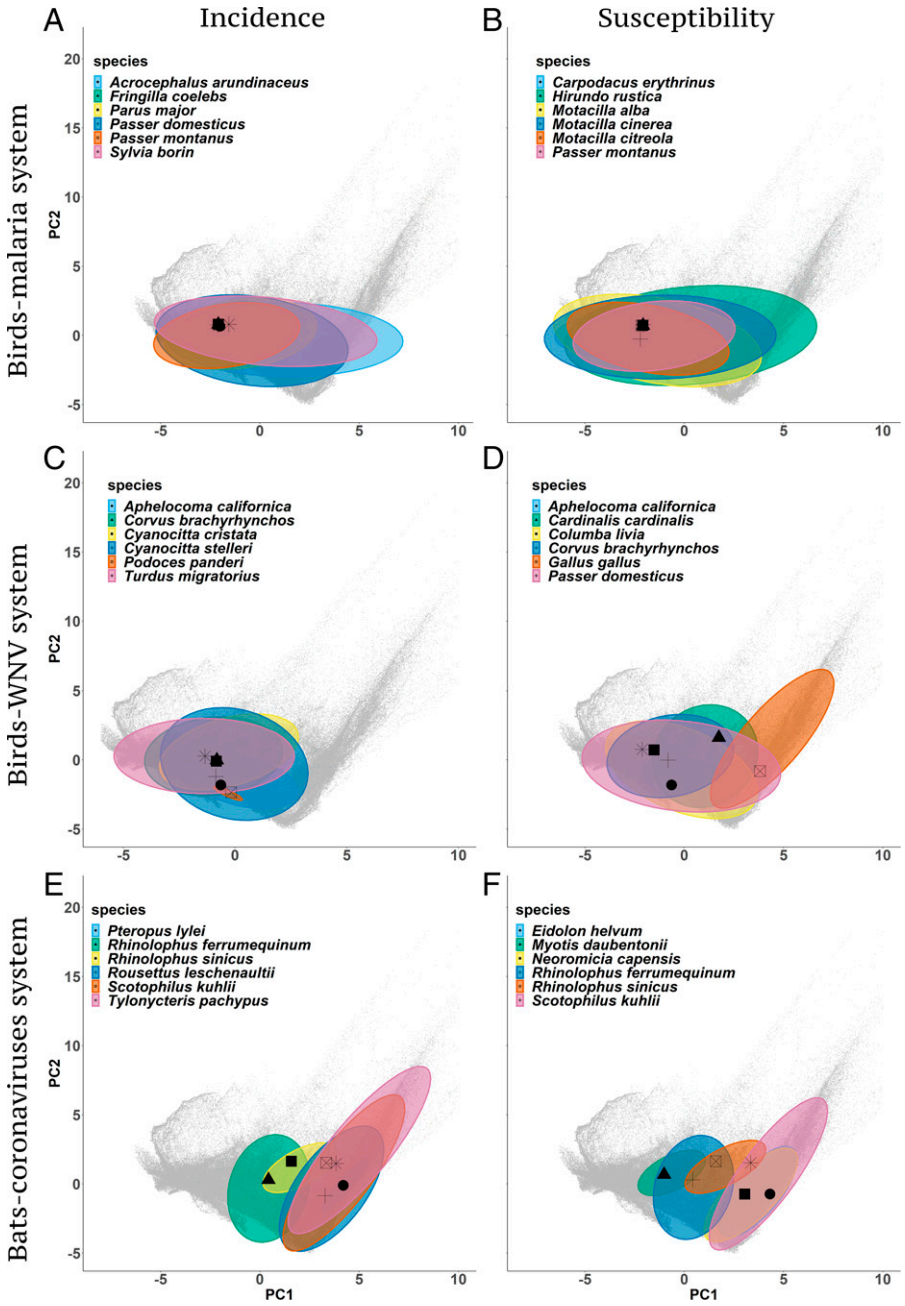

Fig. 5. Environmental space for the three host–pathogen systems. The black shapes represent the centroid of the environmental envelope for each host species. The PC axes summarize the broad temperature (PC1) and precipitation (PC2) conditions at a global scale, where the zero value indicates the average for each axis. (*A*) Environmental envelopes for the top six bird species with highest incidence of *P. relictum*. (*B*) Environmental envelopes for the top six bird species with highest predicted susceptibility to *P. relictum* according to the model. (*C*) Environmental envelopes for the top six bird species with highest incidence of WNV. (*D*) Environmental envelopes for the top six bird species with highest predicted susceptibility to WNV according to the model. (*E*) Environmental envelopes for the top six bat species with highest incidence of coronaviruses. (*F*) Environmental envelopes for the top six bat species with highest predicted susceptibility to coronaviruses according to the model.

the avian malaria system in Hawaii (72), particularly considering the contribution of highly important variables according to this methodological framework.

Finally, it is interesting to note that the well-sampled invasive house sparrow is not among the top 10 most susceptible bird hosts (Table 2; https://doi.org/10.5281/zenodo.6510454). However, there is good agreement between the top 10 bird species with higher incidence and those with higher susceptibility as predicted by our model, where corvids, jays, pigeons, and thrushes are considered highly susceptible to WNV corroborating previous studies (35, 36). In terms of geography, WNV is predicted with a high species richness of susceptible hosts species across North America and western Europe, and lower across central Asia and the territories of Russia, Kazakhstan, Mongolia, and India; yet, some of these regions have no or very low incidence records for this pathogen (e.g., Sundaland and Central Asia), while there is a clear bias toward Western Europe and North America.

Geography was the most important factor determining bat coronavirus susceptibility, which is modulated by the interaction of geography with host phylogeny, and is supported by a study of viral communities in bats and rodents (e.g., ref. 27). Coronaviruses are likely to successfully invade any geographical region of the world that they can reach either via natural dispersal (e.g., bat long-distance migration and host species geographical range

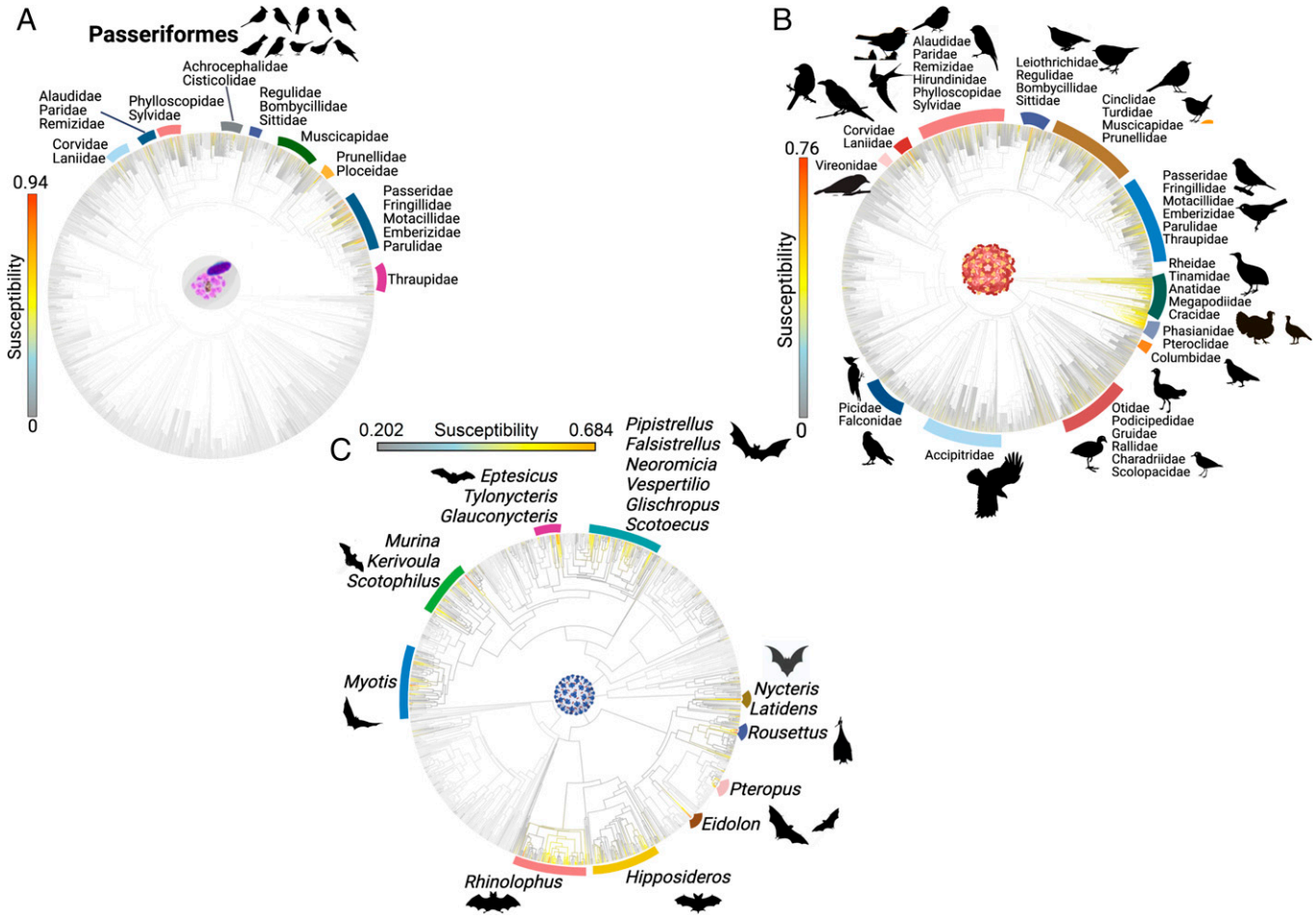

Fig. 6. Susceptibility as a trait mapped onto the phylogeny of the three host–pathogen systems. (*A*) Bird species susceptibility to *P. relictum*. (*B*) Bird species susceptibility to WNV. (*C*) Bat species susceptibility to coronaviruses. The color shapes indicate host families or genera with higher susceptibility to a pathogen.

overlap) or assisted dispersal (e.g., world trade of legal and illegal wild animals). Also, rodents—the most diverse mammal order—can act as an alternative abundant reservoir across the world, aiding in the dispersal and maintenance of bat viruses and other pathogens (27, 45, 73). Phylogenetically conserved host traits play a role in bat virus incidence and dispersal; for instance, roosting behavior in high-density colonies and similar diets increase pathogen transmission (27, 73). In addition to viruses, bats are able to carry other type of Diptera-borne (mainly via flies of the Streblidae and Nycteribiidae families) pathogens such as bacteria (e.g., *Bartonella* spp.) and protozoans (e.g., *Trypanosoma* spp. and *Polychromophilus* spp.) that can become zoonotic in both humans and other animals (73). Given the diverse ecologies and phylogenetic relationships of these parasite systems, and considering that most of the bat fly–parasite interactions are poorly known across different geographical regions (73), we recommend our methodological procedure to be applied in order to discover potential hosts across their geographical distributions. Unlike the cases of avian malaria and WNV, there was correspondence between the top 10 bat species with highest incidence values and those with the highest susceptibility (Table 3; https://doi.org/10.5281/zenodo.6510454). This could be due to a coarser viral taxonomic resolution for this system, which suggests the necessity to conduct a subsequent sensitivity analysis of the current protocol regarding the adequate number of species used for model training and testing.

Our methodological framework can help to direct researchers' attention to including other highly susceptible bat species that may act as reservoirs and represent potential sources (hazards) of emergent diseases (e.g., *Pteropus rodricensis*, *Myotis muricola*, *Rhinolophus beddomei*, and *Laephotis angolensis*; Table 3), particularly in Southeast Asia, tropical Africa, and Europe where our model identified the highest richness of susceptible bat species. Following predictions of host susceptibility for coronaviruses, high-interest regions for sampling correspond to India and parts of the Middle East and Central Asia, where there are currently no incidence records.

The three cases described here demonstrated that variable importance changes across disease systems but that the structure of the community in the context of their phylogenetic, geographic, and environmental signatures is relevant to predict disease susceptibility across (actual or potential) taxa. A subsequent necessary step with this framework would be to apply it to a pathogen capable of crossing the boundaries among several vertebrate clades in order to continue testing the broad applicability of the method.

Although some simple modeling approaches using entropy measures or parasite geographic cooccurrence patterns generate solid predictions of pathogen outbreaks and reemergence potential into the future (e.g., refs. 2, 74), it is of utmost importance to keeping in mind key aspects about the biology of the disease systems (e.g., microhabitat and microclimatic conditions, and diverse functional traits) in order to make appropriate inferences (e.g., ref. 49). One of them is the distinction between the type of transmission (direct or vector-borne; e.g., ref. 2) and type of host (ectothermic vs. endothermic). Thus, the applicability of some methodological frameworks (e.g., refs. 20, 75, 76) as a general

Table 1. Avian malaria incidence and predicted susceptibility

Order	Family	Species	Susceptibility	Incidence
Top 10 species by incidence				
Passeriformes	Passeridae	*P. domesticus*	0.832	188
Passeriformes	Paridae	*Parus major*	0.754	50
Passeriformes	Acrocephalidae	*Acrocephalus arundinaceus*	0.824	23
Passeriformes	Sylviidae	*Sylvia borin*	0.726	19
Passeriformes	Passeridae	*Passer montanus*	0.882	17
Passeriformes	Fringillidae	*Fringilla coelebs*	0.760	16
Passeriformes	Sylviidae	*Sylvia melanocephala*	0.654	16
Passeriformes	Cettiidae	*Cettia cetti*	0.010	12
Passeriformes	Muscicapidae	*Ficedula albicollis*	0.128	12
Passeriformes	Sylviidae	*Sylvia atricapilla*	0.506	12
Top 10 species by susceptibility				
Passeriformes	Passeridae	*P. montanus*	0.882	17
Passeriformes	Motacillidae	*Motacilla citreola*	0.87	NA
Passeriformes	Motacillidae	*Motacilla alba*	0.866	NA
Passeriformes	Motacillidae	*Motacilla cinerea*	0.852	NA
Passeriformes	Fringillidae	*Carpodacus erythrinus*	0.846	4
Passeriformes	Hirundinidae	*H. rustica*	0.836	1
Passeriformes	Passeridae	*P. domesticus*	0.832	188
Passeriformes	Motacillidae	*A. schoenobaenus*	0.832	3
Passeriformes	Fringillidae	*A. godlewskii*	0.832	NA
Passeriformes	Passeridae	*F. montifringilla*	0.832	NA

tool can be difficult and somewhat idiosyncratic, thus the need to have robust and easy-to-use tools that incorporate variables that are relevant in multiple biogeographic contexts, as well as different spatial and temporal scales. Ease in implementation comes at a cost because of oversimplification of the relationship of the variables in each of the disease systems (i.e., as we did here by using a measure of the central tendency of the distances between host–pathogen interactions). Nonetheless, our results confirmed that it is possible to classify host susceptibility with machine learning protocols. Moreover, the variables used for modeling may explain diverse biological phenomena other than susceptibility to a disease, allowing this framework to be used as a means to postulate different hypotheses regarding the biogeography of biotic interactions. These are methodological aspects that could be modified by the user depending on the question and the spatial and temporal scale of the analysis, as well as on the availability of data [e.g., taxonomic distances instead of phylogenetic distances (77) and past or future scenarios (49, 78)].

Here we showed that the three antagonistic systems have a close match to a power law distribution, similar to what Jordano et al. (79) showed for the topology of mutualistic interaction networks. The avian malaria and the coronavirus systems seem to follow a pattern more similar to a truncated power law distribution, however, which was also identified for mutualistic interactions

Table 2. WNV incidence and predicted susceptibility

Order	Family	Species	Susceptibility	Incidence
Top 10 species by incidence				
Passeriformes	Corvidae	*Corvus brachyrhynchos*	0.764	10,014
Galliformes	Phasianidae	*Gallus gallus*	0.640	4,777
Passeriformes	Corvidae	*Aphelocoma californica*	0.746	3,433
Passeriformes	Passeridae	*P. domesticus*	0.668	1,949
Passeriformes	Cardinalidae	*Cardinalis cardinalis*	0.618	1,294
Columbiformes	Columbidae	*Columba livia*	0.684	919
Passeriformes	Corvidae	*Cyanocitta cristata*	0.755	801
Galliformes	Odontophoridae	*Callipepla californica*	0.700	612
Columbiformes	Columbidae	*Zenaida macroura*	0.672	610
Gruiformes	Rallidae	*Fulica atra*	0.664	365
Top 10 species by susceptibility				
Passeriformes	Corvidae	*C. brachyrhynchos*	0.764	10,014
Passeriformes	Corvidae	*C. cristata*	0.755	801
Passeriformes	Corvidae	*A. californica*	0.746	3,433
Passeriformes	Turdidae	*Turdus migratorius*	0.738	319
Passeriformes	Corvidae	*Podoces panderi*	0.73	NA
Passeriformes	Corvidae	*Cyanocitta stelleri*	0.726	207
Passeriformes	Sittidae	*Sitta villosa*	0.718	NA
Passeriformes	Sittidae	*Sitta canadensis*	0.717	1
Galliformes	Odontophoridae	*C. californica*	0.7	612
Passeriformes	Corvidae	*Podoces pleskei*	0.698	NA

Table 3. Coronaviruses incidence and predicted susceptibility

Order	Family	Species	Susceptibility	Incidence
Top 10 species by incidence				
Chiroptera	Pteropodidae	*Eidolon helvum*	0.630	435
Chiroptera	Rhinolophidae	*Rhinolophus sinicus*	0.634	184
Chiroptera	Rhinolophidae	*Rhinolophus ferrumequinum*	0.640	166
Chiroptera	Vespertilionidae	*Scotophilus kuhlii*	0.680	161
Chiroptera	Vespertilionidae	*Myotis daubentonii*	0.632	115
Chiroptera	Vespertilionidae	*Neoromicia capensis*	0.632	110
Chiroptera	Pteropodidae	*Rousettus leschenaultii*	0.665	101
Chiroptera	Pteropodidae	*Pteropus lylei*	0.684	86
Chiroptera	Pteropodidae	*Rousettus aegyptiacus*	0.628	71
Chiroptera	Vespertilionidae	*Tylonycteris pachypus*	0.658	67
Top 10 species by susceptibility				
Chiroptera	Pteropodidae	*P. lylei*	0.684	86
Chiroptera	Vespertilionidae	*S. kuhlii*	0.68	161
Chiroptera	Pteropodidae	*R. leschenaultii*	0.665	101
Chiroptera	Vespertilionidae	*T. pachypus*	0.658	67
Chiroptera	Rhinolophidae	*R. ferrumequinum*	0.64	166
Chiroptera	Rhinolophidae	*R. sinicus*	0.634	184
Chiroptera	Vespertilionidae	*M. daubentonii*	0.632	115
Chiroptera	Vespertilionidae	*N. capensis*	0.632	110
Chiroptera	Pteropodidae	*E. helvum*	0.63	435
Chiroptera	Pteropodidae	*R. aegyptiacus*	0.628	71

(79). This slight nonsignificant departure from a power law may be due to fewer data in the case of avian malaria and due to the taxonomic resolution level for the case of bat coronavirus that was fitted using family-level data. Thus, our study suggests that the use of the power law cumulative density function is a good starting point to estimate the number of incidences that will be necessary for model training; additionally, they seem to be applicable to a broad array of interactions—from mutualistic to antagonistic ones. Importantly, not all recorded incidence cases are necessary to get consistent results after many runs; actually, our models were well trained when using $<20\%$ of the incidence data available. Also, species with higher incidence are not necessarily the species with higher susceptibility. Yet, we suggest that the use of the cumulative density function needs further exploration and implementation, given that, for example, the taxonomic level (e.g., Coronaviridae corresponding to family versus *P. relictum* or WNV which are species) might affect predicted interactions. Model susceptibility results suggest groups of host species that need to be explicitly considered in monitoring efforts, particularly because the mismatch between highly susceptible and high-incidence hosts used in model calibration may indicate a sampling bias toward species more easily captured in the field, as well as the influence of other local-scale factors such as species that are closer to human settlements and interacting more heavily with domestic animals, and the dispersal capacity of host species, among others.

Taking all together, there are several advantages by adopting methodological frameworks such as the one developed in here: 1) the capacity to predict host susceptibility even when there are low or no incidence records, 2) this approximation is statistically robust to sampling biases and easily generalizable to systems other than the ones studied here, and 3) the application to different geographic scales by using a set of relevant variables—here geographic, environmental, and phylogenetic distances at global scales to better understand ecoepidemiological processes. Based on this, future research must include a larger array of interaction systems and a sensitivity analysis of explanatory variables at different spatial and temporal scales (i.e., high-resolution biotic [life history traits and genomics] and abiotic [remote sensing variables] factors).

Materials and Methods

We summarize all the processes from data wrangling to empirical validation of the outputs in Fig. 1.

Host–Pathogen Data. We studied three systems. For the *P. relictum*–birds system, we obtained the data from Malavi database (http://mbio-serv2.mbioekol.lu.se/Malavi/) (80) and filtered the lineages associated with *P. relictum* (SGS1, GRW04, GRW11, LZFUS01, and PHCOL01) (30). For the WNV–birds system we used data from Tolsá et al. (36) considering both serological and molecular prevalence. For the bat coronavirus data we consulted DBatVir (http://www.mgc.ac.cn/DBatVir/) (81). Subsequently, we built datasets for pathogen incidence by species of bird and bat hosts for each of the above mentioned host–pathogen systems; incidence is the number of individuals of a host species recorded infected with a focal parasite species or parasite group.

Incidence Distribution. By incidence we consider all those host species that were infected or exposed (i.e., they had a positive immune reaction to the pathogen) to the pathogen under study. Thus, our incidence data represent presence only or records of positive infected hosts.

We analyzed the statistical hypothesis that the probability distributions of the incidence events in the databases follow a power law distribution (82).

$$p(x) = \frac{\alpha - 1}{x_{min}} \left(\frac{x}{x_{min}} \right)^{-\alpha}. \quad [1]$$

To test this hypothesis we followed Gillespie (83) through poweRlaw R package. In all three cases we accepted the null hypothesis (i.e., data were generated from a power law distribution) and found both x_{min} and α parameters. When we only considered presence data (i.e., incidence), it is likely that we have an excess of host species that have been sampled the most, although they are not necessarily the more abundant ones in the sampling locations.

Yet, in our study cases, ~20 to 40% of hosts would contain 80% of incidence information. Thus, when we build a statistical model with the fitted distribution, the predictions of the model apply to all the host species that follow the power law distribution. This allows us to have a workable classification problem, where the machine learning procedure learns from the data and correctly classifies the information with the set of species of highest incidence. Furthermore, the randomization step guarantees that all host species are considered during model building and calibration (see the step-by-step algorithm in *SI Appendix*, Supplement 1).

Data Sampling. Previous to model calibration with machine learning it is important to identify host species that are considered as "known" versus "unknown" incidence cases. Within known host cases, it is important to select those hosts containing the highest amount of information. To do this, we used the Newton's method (84) to find the optimum set of host species that contain the largest amount of information; the complementary cumulative distribution function (CDF) (82) determines the inflection point for each system, which indicates the highest incidence cases to predict host susceptibility. This process filters the noise that low and unknown incidence species would bring into the modeling procedure as a function of the independent variables.

In this way, we intended that when modeling, the independent variables have a closer relationship with the dataset and the least amount of statistical noise due to interactions because of other unaccounted variables.

Environmental Information. We estimated the environmental distance among host species from their ecological niche centroids based on information from the WorldClim database (https://www.worldclim.org/). To avoid multidimensionality we carried out a principal component analysis (PCA) of the 19 bioclimatic variables and generated new layers from this analysis. We kept the first three PC layers (∼85% of the total variance explained), and for each species we cropped these layers using the International Union for Conservation of Nature (IUCN) shapefiles of their distribution ranges (https://www.iucnredlist.org/resources/spatial-data-download) with the sf R package (85).

In order to understand each climatic variable as a probability density function (PDF) associated with the host geographic distribution (86), we used the kernel density estimation corresponding to the environmental factor across the geographic distribution from the IUCN shapefiles, taking a sample of raster cells within IUCN polygons from PCA environmental layers:

$$p_g(\mathbf{x}) = \frac{1}{k}\sum_{i=1}^{k}\frac{1}{h_x}\frac{1}{\sqrt{2\pi}}e^{-\frac{(\mathbf{x}-\mathbf{x}_i)^2}{2h_x^2}}, \qquad [2]$$

where $p_g(\mathbf{x})$ is the PDF of a species geographical distribution with respect to each environmental factor in a vector $\mathbf{x}$, $\mathbf{x}_i$ is the value of x at each i cell in environmental raster layer, k is the total number of raster cells sampled, and h_x is the bandwidth parameter. Here we adopted a Gaussian kernel (87) and obtained the maximum of the distribution for each variable $r(\mathbf{x}) = max(p_g(\mathbf{x}))$, understanding that this value summarizes the most probable environmental information associated with each species (88).

In this way, each species is associated with a point (i.e., the realized niche centroid) in the environmental space, and it is possible to calculate a distance matrix efficiently for each pair of points in environmental space. An Euclidean matrix $\mathbf{E}$ of environmental distances of all host species was generated. This procedure was performed independently for each host–pathogen system:

$$\mathbf{E} = (e_{ij})$$
$$e_{ij} = |r_i - r_j|, \qquad [3]$$

where e_{ij} is the distance between r_{ij}, the maxima of two $p(\mathbf{x})$ distributions. Finally, because the distances between hosts present a central tendency in most of the cases, we took the average distance in order to summarize the interaction information of each host species in relation to the other host species, collapsing the rows in the matrix by

$$s(e)_n = \frac{\sum_{i\neq j}^{n} e_{ij}}{n}, \qquad [4]$$

and generated a table of environmental distances for n species for each dataset. This approach allowed more efficient computation without losing information associated with each species, and it also allowed the information to be compared between/among species. Because our objective was to describe the environment for each species with the available geographic data, we took the precaution of considering this multivariate probability density function as a continuous function, which permits the method to be applied to understanding the Hutchinsonian niche of the species in order to compare host species to each other (89). To represent the environmental information in a reliable way, an ellipsoid was considered as a hypothetical fundamental niche model for each species (90). To calculate this ellipsoid we used the dataEllipse function of the R package car (91) implemented in the stat_elipse function of the ggplot2 package (92).

Geographical Information. We obtained geographic distribution information for birds and mammals from the IUCN polygons. We first calculated the centroid of the polygon with the largest area for each species to get the geographic information. Subsequently, we calculated the geographical distance between each centroid to generate a geographic distance matrix. Similar to the environmental distance, we obtained a list of distances of one species with respect to the others and calculated the mean distance between the centroids of each pair, getting a table of geographic distances by species expressed as $s(g)_n$ (see above in *Environmental Information* subsection of *Materials and Methods*). We performed all geospatial data manipulation using the sf R package (85).

Phylogenetic Information. We calculated the phylogenetic distance among bird species following Jetz (https://birdtree.org/) (93). In the case of mammals we used Upham tree (94) and calculated a phylogenetic distance matrix with the ape R package (95) given in million years among tips of the tree. Subsequently, we transformed the distance matrix into a list of pairs of species with the distance between them and generated a table per species with the mean phylogenetic distance of one species with respect to the other species. Thus, it is possible to generate a feature that captures the position of a species in the phylogenetic information space. We did this method for each of the phylogenetic trees used (i.e., birds and bats) and expressed them as $s(d)_n$.

Data Modeling. For each host–pathogen assemblage, we created a dataset taking each n host species and the phylogenetic, environmental, and geographic distances calculated above (see subsections *Environmental Information, Geographical Information*, and *Phylogenetic Information*). Based on this dataset, we selected the species considered in the pathogen incidence cutoff (see above in *Data Sampling* subsection of *Materials and Methods*) and labeled these species as susceptible. Subsequently, we took a random sample of species outside of the cutoff set, balancing the sample size with respect to the cutoff set and labeling these species as unknown, thus having a dataset with two susceptibility classes (susceptible and unknown) and three features (environmental, geographic, and phylogenetic distances). We also added the interactions between the independent variables and separated the data into a train set (70%) and a validation set (30%).

For the machine learning procedure, we first performed a comparison of five algorithms for the American mammals–dengue dataset that we previously analyzed (17). We showed that random forest algorithms outperformed other methods (e.g., glmnet and logistic regression; see details of comparisons in Zenodo; https://doi.org/10.5281/zenodo.6510454). Therefore, following Kuhn et al. (96), we generated a set of random forest models optimizing their parameters with a modeling grid (97) and with 10 times 10-fold cross-validation for each sample. We briefly describe the random forest algorithm:

1. For b = 0 to B random variables:
   - Select a bootstrap sample of size N from training data
   - Grow a random forest tree T_b to the bootstrapped data by recursively repeating the following steps for each terminal node of the tree, until the minimum node size n_{min} is reached.
     (a) Select m variables at random from the p variables.
     (b) Pick the best variable/split-point among the m.
     (c) Split the node into two daughter nodes.
2. Output the sample of trees T_{b1}^{B}.

To make a new prediction of susceptible or unknown class, let $\hat{C}_b(x)$ be the class prediction of the bth random forest tree. Then,

$$\hat{C}_{rf}^{B}(x) = \text{majority} \quad \text{vote} \quad \hat{C}_b(x)_1^B. \qquad [5]$$

With the most optimal model we regressed the entire dataset and obtained the probability of susceptibility given the distances and their interactions, also calculating the importance of each variable using the mean decrease impurity (98) implemented internally in the ranger R (99) package. We generated the variable important plots with vip (100) R package. We classified the species as susceptible with a standard probability threshold (i.e., $p(x) > 0.5$),

which is considered acceptable for model accuracy and is useful to select the pool of species denominated as susceptible for the three host–pathogen systems.

Finally, we repeated this procedure 1,000 times with a different sample looking for convergence in the probability calculation, as well as its uncertainty. We then estimated the mean probability of susceptibility for each host species in each host–pathogen assemblage, along with its SE (101).

Susceptibility of Geographical Richness. We assigned the average susceptibility probability of each host species to each shape. Subsequently, we filtered the susceptible species (i.e., with average probability of being susceptible $\hat{p}(x) > 0.5$) and generated a susceptible species richness map that we projected onto the geography for each host–pathogen system using the raster (102) and fasterize (103) R packages.

Statistical Validation of Spatial Patterns. To test our model on geographic space, we placed empirical validation points of where pathogens have been found in the field (i.e., georeferences from field incidence observations; *Host–Pathogen Data*) and applied point intensity tests to statistically analyze the hypothesis that the pattern of observed (i.e., empirical) incidence does not follow a random association with respect to our susceptibility richness map.

We tested as null hypothesis (H_0) that the density of empirical pathogen points is not a function of the richness of susceptible host species and as alternative hypothesis (H_a) that the density of empirical pathogen points depends on the richness of susceptible host species according to the random forest model. We performed this test using the likelihood ratio test for each hypothesis, getting in all cases $p < 0.05$ and rejecting H_0 in all cases. We implemented this analysis with the spatstats R package (104, 105). Additionally, we generated a nonparametric estimate of the intensity of this point process as a function of the richness of those species predicted as susceptible (106). This allowed us to inspect potential unobserved areas (i.e., areas with no records) in geographic space.

Phylogenetic Reconstruction of Susceptibility. The predicted susceptibility of each host–pathogen assemblage was used as a trait to map it onto a phylogeny that was constructed using contmap and fastAnc functions from phytools R package (107).

Data, Materials, and Software Availability. All R scripts and example data are avaliable throught GitHub (https://github.com/alrobles/PNAS-Wildlife-susceptibility-to-infectious-diseases-at-global-scales) (108). RDS files available through Zenodo (https://doi.org/10.5281/zenodo.6510454) (109) contain sufficient information to replicate the analysis for each of the three host–pathogen systems and include initial inputs to calculate the models (the phylogenetic, environmental, and geographic distance tables and the incidence tables), model outputs, accuracy for 1,000 runs, and susceptibility for each system.

ACKNOWLEDGMENTS. Á.L.R.-F. was awarded with the Global Biodiversity Information Facility (GBIF) Young Researchers Award and supported in part by the US NIH (grant 1R01AI151144-01A1). Á.L.R.-F. received scholarship support from Consejo Nacional de Ciencia y Tecnología (CONACYT; 895979). This project was inspired by previous academic exchange with Gregory Gilbert on the 2017 University of California Institute for Mexico and the United States (UC MEXUS)–CONACYT Grant (CN-17-112) "Linking biodiversity dimensions for pest and pathogen risk assessment: The roles of phylogenetic signal, geographic distributions and ecological niches." We especially thank Nathan Upham for his help with reviewing the manuscript. A.L.-N. is a CONACYT Research Fellow. Special thanks to the anonymous reviewers and the editor whose comments helped to improve the quality of this article.

1. P. R. Hosseini *et al.*, Does the impact of biodiversity differ between emerging and endemic pathogens? The need to separate the concepts of hazard and risk. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **372**, 20160129 (2017).
2. T. A. Dallas, C. J. Carlson, T. Poisot, Testing predictability of disease outbreaks with a simple model of pathogen biogeography. *R. Soc. Open Sci.* **6**, 190883 (2019).
3. C. J. Carlson *et al.*, Parasite biodiversity faces extinction and redistribution in a changing climate. *Sci. Adv.* **3**, e1602422 (2017).
4. C. L. Wood *et al.*, Human impacts decouple a fundamental ecological relationship–The positive association between host diversity and parasite diversity. *Glob. Change Biol.* **24**, 3666–3679 (2018).
5. F. R. Adler, C. J. Tanner, *Urban Ecosystems: Ecological Principles for the Built Environment* (Cambridge University Press, 2013).
6. R. Gibb *et al.*, Zoonotic host diversity increases in human-dominated ecosystems. *Nature* **584**, 398–402 (2020).
7. L. Jones-Engel *et al.*, Diverse contexts of zoonotic transmission of simian foamy viruses in Asia. *Emerg. Infect. Dis.* **14**, 1200–1208 (2008).
8. K. F. Smith *et al.*, Global rise in human infectious disease outbreaks. *J. R. Soc. Interface* **11**, 20140950 (2014).
9. F. C. Ferreira-Junior *et al.*, A new pathogen spillover from domestic to wild animals: *Plasmodium juxtanucleare* infects free-living passerines in Brazil. *Parasitology* **145**, 1949–1958 (2018).
10. J. M. Hassell, M. Begon, M. J. Ward, E. M. Fèvre, Urbanization and disease emergence: Dynamics at the wildlife–livestock–human interface. *Trends Ecol. Evol.* **32**, 55–67 (2017).
11. A. Marzal, L. Garcia-Longoria, *The Role of Malaria Parasites in Invasion Biology in Avian Malaria and Related Parasites in the Tropics* (Springer, 2020), pp. 487–512.
12. T. Allen *et al.*, Global hotspots and correlates of emerging zoonotic diseases. *Nat. Commun.* **8**, 1124 (2017).
13. K. J. Olival *et al.*, Host and viral traits predict zoonotic spillover from mammals. *Nature* **546**, 646–650 (2017).
14. C. K. Johnson *et al.*, Global shifts in mammalian population trends reveal key predictors of virus spillover risk. *Proc. R. Soc. B.* **287**, 20192736 (2020).
15. N. Mollentze, D. G. Streicker, Viral zoonotic risk is homogenous among taxonomic orders of mammalian and avian reservoir hosts. *Proc. Natl. Acad. Sci. U.S.A.* **117**, 9423–9430 (2020).
16. A. Uwimana *et al.*, Emergence and clonal expansion of in vitro artemisinin-resistant Plasmodium falciparum kelch13 R561H mutant parasites in Rwanda. *Nat. Med.* **26**, 1602–1608 (2020).
17. Á. L. Robles-Fernández, D. Santiago-Alarcon, A. Lira-Noriega, American mammals susceptibility to dengue according to geographical, environmental, and phylogenetic distances. *Front. Vet. Sci.* **8**, 604560 (2021).
18. A. E. Gorbalenya *et al.*; Coronaviridae Study Group of the International Committee on Taxonomy of Viruses, The species severe acute respiratory syndrome-related coronavirus: Classifying 2019-nCoV and naming it SARS-CoV-2. *Nat. Microbiol.* **5**, 536–544 (2020).
19. R. Poulin, F. Guilhaumon, H. S. Randhawa, J. L. Luque, D. Mouillot, Identifying hotspots of parasite diversity from species–area relationships: Host phylogeny versus host ecology. *Oikos* **120**, 740–747 (2011).
20. G. S. Gilbert, R. Magarey, K. Suiter, C. O. Webb, Evolutionary tools for phytosanitary risk analysis: Phylogenetic signal as a predictor of host range of plant pests and pathogens. *Evol. Appl.* **5**, 869–878 (2012).
21. N. J. Clark, S. M. Clegg, Integrating phylogenetic and ecological distances reveals new insights into parasite host specificity. *Mol. Ecol.* **26**, 3074–3086 (2017).
22. J. Sotomayor-Bonilla *et al.*, Insights into the host specificity of mosquito-borne flaviviruses infecting wild mammals. *EcoHealth* **16**, 726–733 (2019).
23. W. Dáttilo *et al.*, Species-level drivers of mammalian ectoparasite faunas. *J. Anim. Ecol.* **89**, 1754–1765 (2020).
24. K. J. Olival, E. O. Stiner, S. L. Perkins, Detection of *Hepatocystis* sp. in southeast Asian flying foxes (Pteropodidae) using microscopic and molecular methods. *J. Parasitol.* **93**, 1538–1540 (2007).
25. A. Fecchio *et al.*, Avian host composition, local speciation and dispersal drive the regional assembly of avian malaria parasites in South American birds. *Mol. Ecol.* **28**, 2681–2693 (2019).
26. B. R. Krasnov, G. S. Shenbrot, L. van der Mescht, E. M. Warburton, I. S. Khokhlova, Phylogenetic and compositional diversity are governed by different rules: A study of fleas parasitic on small mammals in four biogeographic realms. *Ecography* **42**, 1000–1011 (2019).
27. A. D. Luis *et al.*, Network analysis of host-virus communities in bats and rodents reveals determinants of cross-species transmission. *Ecol. Lett.* **18**, 1153–1162 (2015).
28. G. Valkiunas, C. T. Atkinson, *Introduction to Life Cycles, Taxonomy, Distribution, and Basic Research Techniques in Avian Malaria and Related Parasites in the Tropics* (Springer, 2020), pp. 45–80.
29. E. H. Paxton, M. Laut, J. P. Vetter, S. J. Kendall, Research and management priorities for Hawaiian forest birds. *Condor* **120**, 557–565 (2018).
30. J. Martínez-de la Puente, D. Santiago-Alarcon, V. Palinauskas, S. Bensch, *Plasmodium relictum*. *Trends Parasitol.* (2021), vol. 37, pp. 355–356.
31. V. A. Ellis *et al.*, The global biogeography of avian haemosporidian parasites is characterized by local diversification and intercontinental dispersal. *Parasitology* **146**, 213–219 (2019).
32. A. Fecchio *et al.*, Climate variation influences host specificity in avian malaria parasites. *Ecol. Lett.* **22**, 547–557 (2019).
33. N. J. Clark *et al.*, Climate, host phylogeny and the connectivity of host communities govern regional parasite assembly. *Divers. Distrib.* **24**, 13–23 (2018).
34. J. Sotomayor-Bonilla *et al.*, Survey of mosquito-borne flaviviruses in the Cuitzmala River Basin, Mexico: Do they circulate in rodents and bats? *Trop. Med. Health* **46**, 35 (2018).
35. A. M. Kilpatrick, P. Daszak, M. J. Jones, P. P. Marra, L. D. Kramer, Host heterogeneity dominates West Nile virus transmission. *Proc. R. Soc. B* **273**, 2327–2333 (2006).
36. M. J. Tolsá, G. E. García-Peña, O. Rico-Chávez, B. Roche, G. Suzán, Macroecology of birds potentially susceptible to West Nile virus. *Proc. Biol. Sci.* **285**, 20182178 (2018).
37. K. A. Padgett *et al.*, West Nile virus infection in tree squirrels (Rodentia: Sciuridae) in California, 2004-2005. *Am. J. Trop. Med. Hyg.* **76**, 810–813 (2007).
38. A. T. Peterson, D. A. Vieglais, J. K. Andreasen, Migratory birds modeled as critical transport agents for West Nile Virus in North America. *Vector Borne Zoonotic Dis.* **3**, 27–37 (2003).
39. W. K. Reisen *et al.*, Overwintering of West Nile virus in southern California. *J. Med. Entomol.* **43**, 344–355 (2006).
40. M. C. Wimberly, M. B. Hildreth, S. P. Boyte, E. Lindquist, L. Kightlinger, Ecological niche of the 2003 West Nile virus epidemic in the northern great plains of the United States. *PLoS One* **3**, e3744 (2008).
41. S. J. Anthony *et al.*; PREDICT Consortium, Global patterns in coronavirus diversity. *Virus Evol.* **3**, vex012 (2017).
42. B. Hu *et al.*, Discovery of a rich gene pool of bat SARS-related coronaviruses provides new insights into the origin of SARS coronavirus. *PLoS Pathog.* **13**, e1006698 (2017).
43. H. K. H. Luk, X. Li, J. Fung, S. K. P. Lau, P. C. Y. Woo, Molecular epidemiology, evolution and phylogeny of SARS coronavirus. *Infect. Genet. Evol.* **71**, 21–30 (2019).

44. S. Yuan, S. C. Jiang, Z. L. Li, Analysis of possible intermediate hosts of the new coronavirus SARS-CoV-2. *Front. Vet. Sci.* **7**, 379 (2020).
45. A. D. Luis *et al.*, A comparison of bats and rodents as reservoirs of zoonotic viruses: Are bats special? *Proc. R. Soc. B.* **280**, 20122753 (2013).
46. H. S. Young *et al.*, Declines in large wildlife increase landscape-level prevalence of rodent-borne disease in Africa. *Proc. Natl. Acad. Sci. U.S.A.* **111**, 7036–7041 (2014).
47. H. S. Young *et al.*, Interacting effects of land use and climate on rodent-borne pathogens in central Kenya. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **372**, 20160116 (2017).
48. D. Santiago-Alarcon *et al.*, Parasites in space and time: A case study of haemosporidian spatiotemporal prevalence in urban birds. *Int. J. Parasitol.* **49**, 235–246 (2019).
49. J. M. Cohen, E. L. Sauer, O. Santiago, S. Spencer, J. R. Rohr, Divergent impacts of warming weather on wildlife disease risk across climates. *Science* **370**, eabb1702 (2020).
50. C. Hernández-Lara, P. Carbó-Ramírez, D. Santiago-Alarcon, Effects of land use change (rural-urban) on the diversity and epizootiological parameters of avian Haemosporida in a widespread neotropical bird. *Acta Trop.* **209**, 105542 (2020).
51. M. Q. Wilber, P. T. J. Johnson, C. J. Briggs, Disease hotspots or hot species? Infection dynamics in multi-host metacommunities controlled by species identity, not source location. *Ecol. Lett.* **23**, 1201–1211 (2020).
52. P. T. Johnson, D. L. Preston, J. T. Hoverman, B. E. LaFonte, Host and parasite diversity jointly control disease risk in complex communities. *Proc. Natl. Acad. Sci. U.S.A.* **110**, 16916–16921 (2013).
53. P. T. Johnson, D. L. Preston, J. T. Hoverman, K. L. Richgels, Biodiversity decreases disease through predictable changes in host community competence. *Nature* **494**, 230–233 (2013).
54. W. van Hoesel *et al.*, Management of ecosystems alters vector dynamics and haemosporidian infections. *Sci. Rep.* **9**, 8779 (2019).
55. W. van Hoesel, D. Santiago-Alarcon, A. Marzal, S. C. Renner, Effects of forest structure on the interaction between avian hosts, dipteran vectors and haemosporidian parasites. *BMC Ecol.* **20**, 47 (2020).
56. A. T. Peterson *et al.*, *Ecological Niches and Geographic Distributions (MPB-49)* (Princeton University Press, Princeton, NJ, 2011).
57. P. T. Johnson *et al.*, Habitat heterogeneity drives the host-diversity-begets-parasite-diversity relationship: Evidence from experimental and field studies. *Ecol. Lett.* **19**, 752–761 (2016).
58. O. Hellgren *et al.*, Global phylogeography of the avian malaria pathogen plasmodium relictum based on msp1 allelic diversity. *Ecography* **38**, 842–850 (2015).
59. V. A. Ellis *et al.*, Explaining prevalence, diversity and host specificity in a community of avian haemosporidian parasites. *Oikos* **129**, 1314–1329 (2020).
60. I. Newton, *Speciation and Biogeography of Birds* (Academic Press, United Kingdom, 2003).
61. C. A. Abella-Medrano, S. Ibáñez-Bernal, I. MacGregor-Fors, D. Santiago-Alarcon, Spatiotemporal variation of mosquito diversity (Diptera: Culicidae) at places with different land-use types within a neotropical montane cloud forest matrix. *Parasit. Vectors* **8**, 487 (2015).
62. K. Rodríguez-Hernández *et al.*, Haemosporidian prevalence, parasitaemia and aggregation in relation to avian assemblage life history traits at different elevations. *Int. J. Parasitol.* **51**, 365–378 (2021).
63. M. García-Arroyo, D. Santiago-Alarcon, J. Quesada, I. MacGregor-Fors, Are invasive house sparrows a nuisance for native avifauna when scarce? *Urban Ecosyst.* **23**, 793–802 (2020).
64. A. Marzal *et al.*, Diversity, loss, and gain of malaria parasites in a globally invasive bird. *PLoS One* **6**, e21905 (2011).
65. A. Marzal, L. García-Longoria, J. M. Cárdenas Callirgos, R. N. Sehgal, Invasive avian malaria as an emerging parasitic disease in native birds of Peru. *Biol. Invasions* **17**, 39–45 (2015).
66. D. Bisanzio *et al.*, Evidence for West Nile virus spillover into the squirrel population in Atlanta, Georgia. *Vector Borne Zoonotic Dis.* **15**, 303–310 (2015).
67. M. E. Morales-Betoulle *et al.*; Arbovirus Ecology Work Group, West Nile virus ecology in a tropical ecosystem in Guatemala. *Am. J. Trop. Med. Hyg.* **88**, 116–126 (2013).
68. P. Minias, E. Pikus, L. A. Whittingham, P. O. Dunn, Evolution of copy number at the MHC varies across the avian tree of life. *Genome Biol. Evol.* **11**, 17–28 (2019).
69. D. M. Chisenhall, C. N. Mores, Diversification of West Nile virus in a subtropical region. *Virol. J.* **6**, 106 (2009).
70. D. A. LaPointe, M. L. Goff, C. T. Atkinson, Thermal constraints to the sporogonic development and altitudinal distribution of avian malaria *Plasmodium relictum* in Hawai'i. *J. Parasitol.* **96**, 318–324 (2010).
71. M. S. Shocket *et al.*, Transmission of West Nile and five other temperate mosquito-borne viruses peaks at temperatures between 23°C and 26°C. *eLife* **9**, e58511 (2020).
72. L. B. Fortini, L. R. Kaiser, D. A. LaPointe, Fostering real-time climate adaptation: Analyzing past, current, and forecast temperature to understand the dynamic risk to Hawaiian honeycreepers from avian malaria. *Glob. Ecol. Conserv.* **23**, e01069 (2020).
73. T. Szentiványi, P. Christe, O. Glaizot, Bat flies and their microparasites: Current knowledge and distribution. *Front. Vet. Sci.* **6**, 115 (2019).
74. S. V. Scarpino, G. Petri, On the predictability of infectious disease outbreaks. *Nat. Commun.* **10**, 898 (2019).
75. M. A. Leibold, E. P. Economo, P. Peres-Neto, Metacommunity phylogenetics: Separating the roles of environmental filters and historical biogeography. *Ecol. Lett.* **13**, 1290–1299 (2010).
76. P. T. J. Johnson, D. M. Calhoun, T. Riepe, T. McDevitt-Galles, J. Koprivnikar, Community disassembly and disease: Realistic-but not randomized-biodiversity losses enhance parasite transmission. *Proc. R. Soc. B.* **286**, 20190260 (2019).
77. N. G. Swenson, *Phylogenetic Ecology: A History, Critique, and Remodeling* (University of Chicago Press, 2019).
78. O. Ovaskainen, N. Abrego, *Joint Species Distribution Modelling: With Applications in R, Ecology, Biodiversity and Conservation* (Cambridge University Press, 2020).
79. P. Jordano, J. Bascompte, J. M. Olesen, Invariant properties in coevolutionary networks of plant–animal interactions. *Ecol. Lett.* **6**, 69–81 (2003).
80. S. Bensch, O. Hellgren, J. Pérez-Tris, MalAvi: A public database of malaria parasites and related haemosporidians in avian hosts based on mitochondrial cytochrome b lineages. *Mol. Ecol. Resour.* **9**, 1353–1358 (2009).
81. L. Chen, B. Liu, J. Yang, Q. Jin, DBatVir: The database of bat-associated viruses. *Database (Oxford)* **2014**, bau021 (2014).
82. A. Clauset, C. R. Shalizi, M. E. Newman, Power-law distributions in empirical data. *SIAM Rev.* **51**, 661–703 (2009).
83. C. S. Gillespie, Fitting heavy tailed distributions: The poweRlaw package. *J. Stat. Softw.* **64**, 1–16 (2015).
84. M. P. Deisenroth, A. A. Faisal, C. S. Ong, *Mathematics for Machine Learning* (Cambridge University Press, 2020).
85. E. Pebesma, Simple features for R: Standardized support for spatial vector data. *R J.* **10**, 439–446 (2018).
86. G. Zhang, A. X. Zhu, S. K. Windels, C. Z. Qin, Modelling species habitat suitability from presence-only data using kernel density estimation. *Ecol. Indic.* **93**, 387–396 (2018).
87. B. W. Silverman, *Density Estimation for Statistics and Data Analysis* (CRC Press, 1986), vol. 26.
88. L. Jiménez, J. Soberón, J. A. Christen, D. Soto, On the problem of modeling a fundamental niche from occurrence data. *Ecol. Modell.* **397**, 74–83 (2019).
89. H. Qiao, L. E. Escobar, E. E. Saupe, L. Ji, J. Soberón, A cautionary note on the use of hypervolume kernel density estimators in ecological niche modelling. *Glob. Ecol. Biogeogr.* **26**, 1066–1070 (2017).
90. J. Soberón, A. T. Peterson, What is the shape of the fundamental Grinnellian niche? *Theor. Ecol.* **13**, 105–115 (2020).
91. J. Fox, S. Weisberg, *An R Companion to Applied Regression* (Sage, Thousand Oaks, CA, ed. 3, 2019).
92. H. Wickham, *ggplot2: Elegant Graphics for Data Analysis* (Springer-Verlag, New York, 2016).
93. W. Jetz, G. H. Thomas, J. B. Joy, K. Hartmann, A. O. Mooers, The global diversity of birds in space and time. *Nature* **491**, 444–448 (2012).
94. N. S. Upham, J. A. Esselstyn, W. Jetz, Inferring the mammal tree: Species-level sets of phylogenies for questions in ecology, evolution, and conservation. *PLoS Biol.* **17**, e3000494 (2019).
95. E. Paradis, K. Schliep, ape 5.0: An environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics* **35**, 526–528 (2019).
96. M. Kuhn *et al.*, Building predictive models in r using the caret package. *J. Stat. Softw.* **28**, 1–26 (2008).
97. L. Kjeldgaard, modelgrid: A framework for creating, managing and training multiple caret models (2018), R package version 1.1.1.0.
98. S. Nembrini, I. R. König, M. N. Wright, The revival of the Gini importance? *Bioinformatics* **34**, 3711–3718 (2018).
99. M. N. Wright, A. Ziegler, ranger: A fast implementation of random forests for high dimensional data in C++ and R. arXiv [Preprint] (2015). https://arxiv.org/abs/1508.04409 (Accessed 1 July 2021).
100. B. M. Greenwell, B. C. Boehmke, Variable importance plots–an introduction to the vip package. *R J.* **12**, 343–366 (2020).
101. T. Hastie, R. Tibshirani, J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (Springer Science & Business Media, 2009).
102. R. J. Hijmans, raster: Geographic data analysis and modeling (2022) with RasterData. R Package Version 3.5-21. https://cran.r-project.org/web/packages/raster/index.html. Accessed 1 July 2022.
103. N. Ross, fasterize: Fast polygon to raster conversion (2020). R package version 1.0.3. https://cran.r-project.org/web/packages/fasterize/index.html. Accessed 31 July 2022.
104. A. Baddeley, E. Rubak, R. Turner. *Spatial Point Patterns: Methodology and Applications with R.* (Chapman and Hall/CRC Press, London, UK, 2015).
105. A. E. Gelfand, P. Diggle, P. Guttorp, M. Fuentes, *Handbook of Spatial Statistics* (CRC Press, 2010).
106. A. Baddeley, Y. M. Chang, Y. Song, R. Turner, Nonparametric estimation of the dependence of a spatial point process on spatial covariates. *Stat. Interface* **5**, 221–236 (2012).
107. L. J. Revell, phytools: An R package for phylogenetic comparative biology (and other things). *Methods Ecol. Evol.* **3**, 217–223 (2012).
108. A. L. Robles-Fernandez, Wildlife susceptibility to infectious diseases at global scales. GitHub. https://github.com/alrobles/PNAS-Wildlife-susceptibility-to-infectious-diseases-at-global-scales. Deposited 19 July 2021.
109. A. L. Robles-Fernandez, D. Santiago-Alarcon, A. Lira-Noriega, Predicting wildlife susceptibility to infectious diseases at global scales. Zenodo. https://zenodo.org/record/6510454. Deposited 2 May 2022.