Predicting the tripartite network of mosquito-borne disease

Tad A. Dallas[a,b,†], Sadie J. Ryan[c,d], Ben Bellekom[e], Anna Fagre[f], Rebecca Christofferson[a] and Colin J. Carlson[g,h]

[a] *Department of Biological Sciences, Louisiana State University, Baton Rouge, LA 70802*
[b] *Department of Biological Sciences, University of South Carolina, Columbia, SC, 29208*
[c] *Department of Geography, University of Florida, Gainesville, FL 32611 USA*
[d] *Emerging Pathogens Institute, University of Florida, Gainesville, FL 32611, USA*
[e] *Department of Zoology, 11a Mansfield Road, Oxford OX1 3SZ, UK*
[f] *Department of Microbiology, Immunology, and Pathology, Colorado State University, Fort Collins, CO 80523, US.A.*
[g] *Center for Global Health Science and Security, Georgetown University Medical Center, Washington, D.C. 20007*
[h] *Department of Microbiology and Immunology, Georgetown University Medical Center, Washington, D.C. 20007*
[†] *Corresponding author: tad.a.dallas@gmail.com*

Running title: A tripartite model of vector-borne disease

Author contributions: TAD, CJC, and SJR conceived the study. TAD and CJC performed the analysis. All authors contributed to manuscript writing.

Acknowledgements: Lauren Holian provided comments and discussion on an earlier version of this work. This work has been supported by funding to the Viral Emergence Research Initiative (VERENA) consortium, including a grant from the U.S. National Science Foundation (NSF BII 2021909).

Conflict of interest: The authors have no conflicts of interest to declare.

Keywords: Arbovirus, Host-virus, Link prediction, Networks, Vector-borne

Predicting the tripartite network of mosquito-borne disease

Abstract

The potential for a pathogen to infect a host is mediated by traits of both the host and pathogen, as well as the complex interactions between them. Arthropod-borne viruses (arboviruses) require an intermediate arthropod vector, which introduces an additional layer of compatibility filters. Existing computational models for the prediction of host-virus networks rarely incorporate the unique aspects of vector transmission, instead treating vector biology as a hidden, unobserved layer. Here, we explore two possible extensions to existing approaches, to address this nuance: first, we added vector traits into predictions of the bipartite host-virus network; and second, we used host, vector, and virus traits to predict the tripartite host-vector-virus network. We tested both approaches on the most thoroughly characterized group of arboviruses; mosquito-borne flaviviruses of mammals, including dengue, yellow fever, and Zika virus. Using host-virus models, we find that the inclusion of vector traits may improve inference in some cases, while viral traits proved to be the most important for model performance. Further, we found that it was possible to predict full life cycles (host-vector-virus links), but the model only showed fair performance, and was heavily influenced by the geographic bias of component input datasets (especially the dipteran biting data). Both approaches are interesting avenues for further model development, but our results keenly underscore a need to collect more comprehensive datasets to characterize arbovirus ecology, across a wide geographic scope, especially outside of North America, and to better identify molecular traits that underpin host-vector-virus interactions.

Introduction

Emerging viruses continue to pose a threat to human and wildlife populations [1]. A growing set of computational tools have explored viral dynamics in the context of species interaction networks using a set of tools called *link prediction models*. Typically, these represent hosts and viruses as a bipartite network of either known interactions (that occur in nature [2, 3]) or all possible interactions (including, for example, experimental infections [4]), with both represented as links in the network [5]. Host-virus link prediction models are predominantly trained on the genomic, immunological, morphological, and ecological traits of hosts and viruses (e.g., [6, 7]), while some approaches also leverage information on the latent structure of the network instead of, or in addition to, these traits [8, 9]. The objective of these modeling exercises is to learn about the underlying biology, explain and reproduce patterns found in nature, and anticipate what future dynamics of viral emergence could look like. For example, many models use networks to understand why some viruses can infect humans but others cannot, with the objective of identifying animal viruses that could someday infect humans for the first time. In most cases, these models assume that any given "link" between a host and a virus could represent a self-contained transmission cycle (though not necessarily *onwards* transmission, e.g., West Nile virus in humans and horses [10]).

Vector-borne disease (VBD) transmission substantially complicates this conceptual framework. Vector-borne viruses require an additional species–usually an arthropod (hence arthropod-borne viruses, or *arboviruses*)–to move them between hosts, which adds complexity into their ecology, epidemiology, and evolution. For example, in the case of arboviruses, the presence of both virus and suitable hosts is not necessarily sufficient for transmission, and the presence or absence of suitable vectors (e.g., their geographic distributions or host preferences) may be a latent variable in ecological datasets [11]. Moreover, the "compatibility filters" that can be inferred from the host-virus network will be incomplete, as models will miss both the molecular and physiological determinants of vector-virus compatibility (i.e., vector competence) and the behavioral and ecological determinants

of vector-host compatibility (i.e., biting preferences, in the case of blood-feeding arthropods). If vectors are entirely omitted from the inference process, a model might therefore reach spurious conclusions about whether a given host and virus are incompatible based on their biology, or otherwise miss key drivers of network structure; for example, arboviruses have been shown repeatedly to have a higher-than-expected host breadth [12].

No one canonical approach exists to address vector transmission in link prediction studies. Vector transmission could be described as a binary trait of viruses, which may help make some distinctions (e.g., separating the ecology of mosquito-borne and tick-borne flaviviruses from counterparts like hepatitis C), but leaves much to be desired in terms of information content (e.g., not distinguishing the tick- and mosquito-borne flaviviruses). The possibility of incorporating more detailed information on vector-borne transmission into these models has been under-explored, likely because arboviruses are usually seen as a complicated exception to existing datasets, rather than a feature with significant impacts on network structure. Incorporating traits characterizing the life cycle of arboviruses might improve model performance, given that virus traits are often sparser than host traits, and their interactions usually have non-additive but positive effects on model performance. However, adding sparse traits that only describe some of the viruses in the network could also reduce accuracy if the network includes a mix of vector-borne and directly-transmitted viruses.

Alternately, vectors could be added directly into the network as an additional layer of nodes (Figure 1). While previous work has predicted vector-virus networks [13], none have predicted *host-vector-virus* networks. Existing network models have been used to predict undetected links in *tripartite* networks [14], but this has yet to be explored for ecological networks. This approach would be much more informative than the bipartite form, but also requires difficult-to-obtain data: sylvatic VBD cycles tend to be characterized one at a time in scientific literature (e.g., "*Culex quinquefasciatus* vectors West Nile virus in house finches"). While available datasets could be used to reconstruct these cycles from each of their component

parts (biting preferences, vector competence, and host-virus compatibility), to our knowledge, this has not previously been explored in predictive work.

To address this, we developed two new approaches and tested them on mosquito-borne flaviviruses, a well-studied group that includes important zoonoses like dengue, West Nile, yellow fever, and Zika viruses. Through a synthesis of existing data sources, we combined data on mammal-virus associations [12], vector-flavivirus associations [13], and diptera-mammal biting preferences [15]. We combined these data into one mammal-mosquito-flavivirus network, which can also be reduced down to a mammal-flavivirus network where viruses' mosquito communities are represented as node metadata. Using boosted regression trees (BRT; a machine learning method popular in ecological modeling, also sometimes called gradient boosting machines), we tested two approaches to predicting vector-borne transmission as an aspect of the host-virus network. First, we predicted the mammal-flavivirus network using every possible combination of host, vector, and virus traits, as metadata for any given host-virus association, assuming that additional data layers would enhance model performance. This was generally shown to be true, although the combination of host and vector trait data was not informative compared to the incorporation of viral trait data. Second, we developed a tripartite model of vector-borne disease transmission, in which each link represents a known host-vector-virus link and attempted to predict those complete cycles using traits of hosts, mosquito vectors, and viruses. We found that these models performed more poorly on average, but that they were able to make better than random predictions, including some of relevance to arboviral ecology and human health.

Methods

Host, vector, and virus data Host-virus interaction data were obtained from the CLOVER database [16], a manually- and programmatically-curated database of host-virus associations built by reconciling four disparate datasets (the Host-

Parasite Phylogeny Project, or HP3 [12]; the Global Mammal Parasite Database v2.0 [17]; the Enhanced Infectious Disease Database [18]; and an unnamed dataset curated by Shaw *et al.* [19]). We used CLOVER release 0.1.2, which includes data on 5,477 known interactions between 831 viruses of 1,085 mammal species. These data have been carefully cleaned for taxonomic quality control and include detailed metadata on interaction evidence. These data are also part of a larger open database called The Global Virome in One Network (VIRION), the largest open atlas of vertebrate-virus associations [20]. Although more data is available from this source, we restricted our analysis to the manually-curated data to prevent inclusion of spurious interactions.

Vector-virus association data were taken from a previous study that aimed to predict the mosquito-flavivirus network. [13] These data include 334 associations between 180 mosquito species and 37 flaviviruses. Host-vector association data were taken from a recent study of dipteran biting networks [15]. These data describe 1744 associations between 255 biting dipteran species and 214 hosts (including 67 mammals). Trait data for hosts, vectors, and viruses were assembled from published sources. Thirty-three traits on mosquito life history, ecology, and geography and 22 traits on viral features, were taken from the Evans *et al.* study of the mosquito-flavivirus network [13]. Finally, we used a total of 18 traits on mammal life history, ecology, and morphology from the PanTHERIA database [21].

Modeling approach Boosted regression tree (BRT) models were used to model host-virus and host-vector-virus associations. BRT models have previously been used to model species distributions [22], predict associations in bipartite networks [23, 24, 25, 5], and in other conservation and management settings e.g., [26]. Much of the diversity of applications can be attributed in part to the allowance for nonlinear responses and variable interactions in BRT models. Since the regression tree is hierarchical, "upstream" splits based on one variable influence "downstream" splits, which automatically models variable interactions. Further, the process of boosting enhances learning on complex data, as the process produces many regres-

sion trees with a small number of splits, each of these "weak learners" iteratively build on previous trees to account for the remaining variation. This approach removes the need to partition variance among submodels, as the goal is not to examine the components of variance explained, but to assess overall model performance with the inclusion or exclusion of particular variable sets. Models were trained in the *R* statistical programming language [27] using the *gbm* package [28].

Model 1: Modeling mammal-virus associations as a bipartite network We used the mammal and virus trait data as described above. However, mosquito vector "traits" were created by calculating the number of mosquito species in a given genus which were demonstrated to transmit a particular flavivirus [13]. This is because each host-virus association could be transmitted by any number of mosquito species, creating a range of trait values that may be less informative than simply knowing breadth and composition of the vector community. This resulted in a total of 19 mosquito vector covariates, ranging in value from 0 to 22 species. We removed covariates with less than 25% data coverage, resulting in 13 host traits, 19 mosquito covariates (as virus traits), and 17 virus traits.

The data were split into 80% training and 20% testing sets, where model performance was assessed on the 20% test set. A total of 20 models per covariate group were fit in order to account for the random train/test split. These same 20 train/test divisions were used across the different covariate models, as we trained every possible combination of host, vector, and virus trait data to predict host-virus associations. Together, this resulted in a dataset that allows the estimation of the relative influence of host traits, viral traits, and vector community data on resulting mammal-virus associations. We sampled background data by randomly combining host and virus species, resulting in 25% known positive associations and 75% background data.

We subset these data in two different ways, to explore how vector data may improve prediction of 1) flaviviruses for which we have some vector data (235 known host-virus associations) and 2) all vector-borne viruses (3016 host-virus

associations). This breakdown corresponds to data subsets of 1) only mosquito-borne flaviviruses present in [13] and 2) all viruses that were recorded as vector-borne (or unknown) in the Clover data [16]. We present the flavivirus-specific results here, which are qualitatively similar to the more general models for all vector-borne viruses, which are in the Supplemental Materials.

Model 2: Modeling mammal-mosquito-virus associations as a tripartite network Using the same data resource as used above on host-virus associations, we now considered the identity of the mosquito vector species, and the association between the vector and virus [13], and the feeding association between mosquito vector and mammal species [15]. While host and virus traits were largely the same as considered above, the mosquito vector traits consisted of a set of 33 mosquito vector traits from [13]. Host and virus traits must have 75% of data coverage – the same as in *Model 1* – to be included in this analysis. This resulted in 8 host traits, 29 vector traits, and 16 virus traits. A tripartite link – detailing the full host-vector-virus cycle – was only considered if there were all three associations; host-vector association, vector-virus association, and host-virus association. This creates a situation where a host and vector species may interact, and that vector may be infected by a virus, but this is not a confirmed link if there is no evidence that the host is infected by the virus.

A total of 135 full tripartite links were documented. We sampled background data by randomly combining host, vector, and virus species and then adding enough unique host-vector-virus background points to have 50% true tripartite links and 50% background data. Models were trained in the same manner as in `Model 1`.

Assessing model performance Model performance was quantified using two measures; accuracy and the area under the receiver operating characteristic (AUC). Accuracy was defined as the correctly estimated positives (true positives) and negatives (true negatives) over all the predictions, capturing the fraction of times the model correctly classified host-virus associations in the holdout data. Accuracy is

bounded between 0 and 1, where larger values correspond to higher model performance. AUC is a widely used metric of model discrimination that captures the ability of the classifier to rank positive instances higher than negative instances. AUC is bounded between 0 and 1, where a random model will perform with AUC of 0.5 on average, and values closer to 1 indicate higher model performance.

Data and code availability *R* code and data to reproduce the analyses is available on figshare at
`https://doi.org/10.6084/m9.figshare.17033309`.

Results

Model 1: The mammal-virus models Models trained only on host ($AUC = 0.57$) or vector ($AUC = 0.46$) traits consistently performed poorly at the task of host-virus link prediction (Figure 2), though the viral trait model performed well ($AUC = 0.95$). Generally, combinations of predictor features led to improved model performance. The full model including host, vector, and virus traits performed extremely well ($AUC = 0.96$). However, both the host-virus and vector-virus traits only models also performed extremely well (performance differences among these models were essentially indistinguishable; Figure 2). The inclusion of viral traits seems to have been particularly important; for comparison, the model using host and vector traits to predict host-virus associations barely performed better than random ($AUC = 0.59$).

Variables important for predicting host-virus associations were generally conserved across submodels considering all combinations of host, vector, and virus traits (Figure 3). In the full bipartite model, the most informative variable was whether a virus was found in the Pacific region (likely a proxy for Zika virus, which spread through Pacific islands preceding the epidemic in the Americas). Other important characteristics predictive of host-virus associations in bipartite models including virus traits were disease severity, genome length, year of virus

isolation, if the virus is found in Africa or Australia, and viral clade. In models that omitted virus traits, the top predictors represented host allometry (body mass and metabolic rate, an unsurprising axis of variation) and *Culex* association, which likely captures a latent split between some bird-reservoired viruses (e.g., West Nile virus) and primate-reservoired ones (e.g., dengue and Zika virus).

Overall, our results suggest that models learned from vector trait data, particularly in the full model, where the contribution of each individual variable is more diffuse. However, our findings also indicate that the inclusion of vector data only minimally improved performance after data on hosts and viruses was already available. As host-virus models are usually trained only on host and virus trait data, our findings suggest that the incorporation of vector data into a host-virus model is an imperfect way to explore the role of vectors in structuring the host-virus network. However, this also suggests that improved arthropod trait data could improve model performance, and thus the importance of the vector cannot be overlooked.

Finally, we investigated whether including vector trait data would improve performance even if only available for a subset of data informing the network. To test this, we trained the model on a network that included all the arboviruses present in the CLOVER dataset, even though viral trait data and vector associations were only known for flaviviruses. We found that the model using just host and virus traits performed substantially worse here ($AUC = 0.70$) than the flavivirus-only model with those traits ($AUC = 0.95$). We found that the best performing models were those that used vector and virus traits ($AUC = 0.98$) and those that included host, vector, and virus traits ($AUC = 0.99$; Figure 2). We suggest that this finding indicates that adding data on the vector aspect of transmission may be useful even when it only covers a subset of species in the network.

Model 2: The tripartite model Models trained on tripartite (i.e., host-vector-virus) associations had moderate explanatory power (mean $AUC = 0.64$ (0.065); mean Accuracy = 0.66 (0.046) out of 100 models trained on random subsets). This

lower model performance could simply be due to the smaller amount of data used for training (recall that only 135 full tripartite links were known), or the imbalance between the number of potential full tripartite links given host, vector, and virus diversity, and the small number of realized links (see the small number of red links in Figure 4). Although the model's performance was only fair, we found that the model still predicted higher suitability for tripartite links where one or two of the three possible components were confirmed (Figure 5), even though these would be recorded as a "0" outcome variable the same as if none of them were known. We suggest that this indicates the model was identifying and reproducing real biological signals of compatibility.

The top nine covariates to predicting tripartite (i.e., host-vector-virus) associations were host ($n = 5$) or virus ($n = 4$) traits (Figure 6). The top predictors mostly reflected the geography of transmission (host geographic range size, virus transmission in Asia, vector presence in Africa), the life history of the host (age at first birth, lifespan, weaning age, and neonate body mass), and aspects of viral transmission (genome length and transmission by non-mosquito arthropods).

The predictions made by the tripartite model suggest the model may be able to recover interesting or important biologically-plausible interactions. Both the top predicted "undiscovered" human-mosquito-virus links (Table 1) and mammal-mosquito-virus links (Table 2) heavily over-represent a small number of viruses, in particular Wesselbron virus and West Nile virus. This is driven by the existing level of sampling in the data: West Nile has the greatest number of known hosts ($n = 103$ species) and mosquito vectors ($n = 51$); Wesselbron has the second highest number of vectors ($n = 41$), though many fewer hosts ($n = 11$; ranked #13). This "rich-get-richer" has been previously debated as a strength or weakness for link prediction models; it may be that models are identifying a genuine biological signal of generality (which is known to be true for these viruses), but they may also be recapitulating sampling bias [5, 29] and underpredicting link probabilities for undersampled species. Indeed, the richness of flavivirus data available to us in this study is likely largely due to a discovery and data synthesis bump in the wake

of the Zika virus epidemic in the Americas. The mammal-mosquito-virus predictions also contain a visible signal of geographic bias: most of the top predictions either involve agricultural species (pigs, *Sus scrofa*; cows, *Bos taurus*; or sheep, *Ovis aries*), synanthropic species (black rats, *Rattus rattus*), or charismatic North American species (the opossum, *Didelphis virginiana*; the raccoon, *Procyon lotor*; the white-tailed deer, *Odocoileus virginianus*). These likely reflect a compounded bias between the host-virus association data and the biting data, the latter of which is particularly limited to North American and European species.

Despite the signal of data bias in these predictions, the models reveal several predictions of biological interest. For example, *Anopheles hyrcanus* is predicted as a possible vector of Kokobera virus in humans. The virus was implicated in an outbreak of acute polyarticular illness in Australia in the 1980s based on serology, but it remains poorly understood [30]. The virus was first isolated from *Culex annulirostris*, which also vectors Japanese encephalitis virus and a handful of others; *An. hyrcanus* is a European and Asian mosquito only currently known to vector Japanese encephalitis virus. Similarly, the model predicts that *Culex tritaeniorhynchus* – the main vector of Japanese encephalitis virus, found in southeast Asia – could transmit Murray Valley encephalitis virus in wallabies (*Macropus agilis*). Neither the Australian virus nor the host have been recorded in association with this vector, but as of 2021, the mosquito has been detected in Australia [31], indicating the possibility that this interaction could now emerge.

Discussion

In this study, we considered two approaches to incorporate arboviral life cycles into link prediction models of the mammal-flavivirus network. First, we used a host-virus (bipartite) framework, and assessed the relative influence of including different trait covariates. We found that viral traits were the strongest contributor to model performance, and the incorporation of host and vector traits into the bipartite models did little to improve model performance. Second, we explored how

these models could be extended to predict the entire host-vector-virus (tripartite) network. This framing is both inherently more complex than the host-virus predictive problem, and is massively limited by the availability of training data, but appears promising for future development.

Neither of these approaches provided a complete solution to the host-vector-virus prediction problem, though their limitations differ slightly, with different implications for next steps. Adding vector community data to the bipartite (host-virus) models may be useful where data allow, but may be less important when more detailed, biologically meaningful viral trait data are available. Compared to synthetic datasets of animal ecology, life history, and morphology, only a handful of viral traits (e.g., genome length or disease severity) are available in a standardized format, to the point that viral host range is itself often used as a viral trait (e.g., our "primate" or "bird" traits, or "host breadth" (see Table S1)). Recently, some studies have begun to use immunogenetic or genome composition variables to characterize host and virus compatibility more directly [32, 33, 34, 35, 36, 37]; comparable features for vectors are not yet available or tested in this framework. Shifting towards these kinds of predictors could help models identify more meaningful signals of virus-animal compatibility, and proportionally reduce the signal of bias in predictions.

In contrast, directly modeling the host-vector-virus tripartite network addresses the nuance of vector transmission head-on, but this problem is more severely data limited. As a result, these predictions are very visibly influenced by the geographic and taxonomic bias in the component datasets. However, these data limitations can be addressed by investment in future work characterizing arboviral life cycles in understudied areas [38]. Vector-virus combinations can be tested in the laboratory, including in model-experiment feedback designs that leverage existing predictions (e.g., [13]) much like model-guided fieldwork can be used to optimize viral discovery [25]. Similarly, further investigation of mosquito biting behavior will help resolve the host-vector component [15], highlighting the need for "basic" natural history research even on mosquitoes that are not known to be primary vectors of human

disease.

Our study is the first to attempt modeling the entire tripartite host-vector-virus network. This is a clear knowledge gap in existing approaches to modeling the host-virus network: identifying a suitable host-pathogen association that has no shared vector may not accurately estimate spillover risk. This may be particularly relevant to efforts to identify viruses with undiscovered zoonotic potential, as the presence or absence of human-biting mosquitoes will be a key contributor to their emergence risk [39]. Similarly, the tripartite framework can provide useful insights into the establishment of sylvatic cycles in interepidemic periods or upon expansion into new geographic areas. The ability of arboviruses to persist in non-human hosts may determine whether an epidemic ends as immunity grows (like Zika virus in the Americas, which was primarily transmitted human-to-human by *Aedes aegypti* and *Ae. albopictus*) or instead becomes a regular occurrence (e.g., yellow fever in the Americas, which is maintained by *Haemogogus spp.* and *Sabethes spp.* in non-human primates, between human epidemics driven by *Aedes aegypti*). These are likely to be particularly important nuances as arboviruses continue to spread around an increasingly globalized world in a changing climate [40, 41, 42, 43]

The broader question of "how should we model multi-layer ecological interaction networks" is also one that is likely to have broader implications in computational ecology. For example, there are other cases where researchers are interested the traits that structure tripartite networks, such as bat-bat fly-pathogen networks or plant-pest-parasitoid networks. Multilayer networks are also a topic of increasing interest in network science and mathematics, which will likely open doors for more advanced predictive approaches than the extensions we propose here. This is therefore a promising space for the development of future models, particularly if approached through the lens of iterative validation and data collection [25].

References

[1] Jones, K. E., Patel, N. G., Levy, M. A., Storeygard, A., Balk, D., Gittleman, J. L. & Daszak, P., 2008 Global trends in emerging infectious diseases. *Nature* **451**, 990–993.

[2] Kading, R. C., Kityo, R. M., Mossel, E. C., Borland, E. M., Nakayiki, T., Nalikka, B., Nyakarahuka, L., Ledermann, J. P., Panella, N. A., Gilbert, A. T. *et al.*, 2018 Neutralizing antibodies against flaviviruses, babanki virus, and rift valley fever virus in ugandan bats. *Infection Ecology & Epidemiology* **8**, 1439215.

[3] Fagre, A. C., Lewis, J., Miller, M. R., Mossel, E. C., Lutwama, J. J., Nyakarahuka, L., Nakayiki, T., Kityo, R., Nalikka, B., Towner, J. S. *et al.*, 2021 Subgenomic flavivirus rna (sfrna) associated with asian lineage zika virus identified in three species of ugandan bats (family pteropodidae). *Scientific Reports* **11**, 1–8.

[4] Malmlov, A., Bantle, C., Aboellail, T., Wagner, K., Campbell, C. L., Eckley, M., Chotiwan, N., Gullberg, R. C., Perera, R., Tjalkens, R. *et al.*, 2019 Experimental zika virus infection of jamaican fruit bats (artibeus jamaicensis) and possible entry of virus into brain via activated microglial cells. *PLoS Neglected Tropical Diseases* **13**, e0007071.

[5] Albery, G. *et al.*, 2021 The science of the host-virus network. *Nature Microbiology* p. in press.

[6] Dallas, T., Park, A. W. & Drake, J. M., 2017 Predicting cryptic links in host-parasite networks. *PLoS computational biology* **13**, e1005557.

[7] Elmasri, M., Farrell, M. J., Davies, T. J., Stephens, D. A. *et al.*, 2020 A hierarchical bayesian model for predicting ecological interactions using scaled evolutionary relationships. *Annals of Applied Statistics* **14**, 221–240.

[8] Ghasemian, A., Hosseinmardi, H., Galstyan, A., Airoldi, E. M. & Clauset, A., 2020 Stacking models for nearly optimal link prediction in complex networks. *Proceedings of the National Academy of Sciences* **117**, 23393–23400.

[9] Becker, D. J., Albery, G. F., Kessler, M. K., Lunn, T. J., Falvo, C. A., Czirják, G. Á., Martin, L. B. & Plowright, R. K., 2020 Macroimmunology: The drivers and consequences of spatial patterns in wildlife immune defence. *Journal of Animal Ecology* **89**, 972–995.

[10] Bowen, R. A. & Nemeth, N. M., 2007 Experimental infections with west nile virus. *Current Opinion in Infectious Diseases* **20**, 293–297.

[11] Huang, Y.-J. S., Higgs, S. & Vanlandingham, D. L., 2019 Arbovirus-mosquito vector-host interactions and the impact on transmission and disease pathogenesis of arboviruses. *Frontiers in microbiology* **10**, 22.

[12] Olival, K. J., Hosseini, P. R., Zambrana-Torrelio, C., Ross, N., Bogich, T. L. & Daszak, P., 2017 Host and viral traits predict zoonotic spillover from mammals. *Nature* **546**, 646–650.

[13] Evans, M. V., Dallas, T. A., Han, B. A., Murdock, C. C. & Drake, J. M., 2017 Data-driven identification of potential zika virus vectors. *Elife* **6**, e22053.

[14] Chen, C., Grady, S. K., Ellingson, S. R. & Langston, M. A., 2021 Gene-disease-drug link prediction using tripartite graphs. In *Proceedings of the 12th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics*, pp. 1–1.

[15] Bellekom, B., Hackett, T. D. & Lewis, O. T., 2021 A network perspective on the vectoring of human disease. *Trends in Parasitology* pp. S1471–4922.

[16] Gibb, R., Albery, G. F., Becker, D. J., Brierley, L., Connor, R., Dallas, T. A., Eskew, E. A., Farrell, M. J., Rasmussen, A. L., Ryan, S. J. *et al.*, 2021 Data proliferation, reconciliation, and synthesis in viral ecology. *BioScience* p. in press.

[17] Stephens, P., Pappalardo, P., Huang, S., Byers, J., Farrell, M., Gehman, A., Ghai, R., Haas, S., Han, B., Park, A. *et al.*, 2017 Global mammal parasite database version 2.0. *Ecology* **98**, 1476–1476.

[18] Wardeh, M., Risley, C., McIntyre, M. K., Setzkorn, C. & Baylis, M., 2015 Database of host-pathogen and related species interactions, and their global distribution. *Scientific data* **2**, 1–11.

[19] Shaw, L. P., Wang, A. D., Dylus, D., Meier, M., Pogacnik, G., Dessimoz, C. & Balloux, F., 2020 The phylogenetic range of bacterial and viral pathogens of vertebrates. *Molecular ecology* **29**, 3361–3379.

[20] Carlson, C. J., Gibb, R. J., Albery, G. F., Brierley, L., Connor, R., Dallas, T., Eskew, E. A., Fagre, A. C., Farrell, M. J., Frank, H. K. *et al.*, 2021 The global virome in one network (virion): an atlas of vertebrate-virus associations. *bioRxiv* .

[21] Jones, K. E., Bielby, J., Cardillo, M., Fritz, S. A., O'Dell, J., Orme, C. D. L., Safi, K., Sechrest, W., Boakes, E. H., Carbone, C. *et al.*, 2009 PanTHERIA: a species-level database of life history, ecology, and geography of extant and recently extinct mammals. *Ecology* **90**, 2648–2648.

[22] Elith, J., Leathwick, J. R. & Hastie, T., 2008 A working guide to boosted regression trees. *Journal of Animal Ecology* **77**, 802–813.

[23] Dallas, T. A., Han, B. A., Nunn, C. L., Park, A. W., Stephens, P. R. & Drake, J. M., 2019 Host traits associated with species roles in parasite sharing networks. *Oikos* **128**, 23–32.

[24] Dallas, T. A. & Becker, D. J., 2021 Taxonomic resolution affects host- parasite association model performance. *Parasitology* **148**, 584–590.

[25] Becker, D. J., Albery, G. F., Sjodin, A. R., Poisot, T., Bergner, L. M., Dallas, T. A., Eskew, E. A., Farrell, M. J., Guth, S., Han, B. A. *et al.*, 2021 Optimizing predictive models to prioritize viral discovery in zoonotic reservoirs. *The Lancet Microbe* p. in press.

[26] Soykan, C. U., Eguchi, T., Kohin, S. & Dewar, H., 2014 Prediction of fishing effort distributions using boosted regression trees. *Ecological Applications* **24**, 71–83.

[27] R Core Team, 2020 *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

[28] Greenwell, B., Boehmke, B., Cunningham, J. & Developers, G., 2020 *gbm: Generalized Boosted Regression Models*. R package version 2.1.8.

[29] Gibb, R., Albery, G. F., Mollentze, N. F., Eskew, E. A., Brierley, L., Ryan, S. J., Seifert, S. N. & Carlson, C. J., 2021 Mammal virus diversity estimates are unstable due to accelerating discovery effort. *bioRxiv* .

[30] Boughton, C. R., Hawkes, R. A. & Nairn, H. M., 1986 Illness caused by a kokobera-like virus in south-eastern australia. *Medical journal of Australia* **145**, 90–92.

[31] Lessard, B. D., Kurucz, N., Rodriguez, J., Carter, J. & Hardy, C. M., 2021 Detection of the japanese encephalitis vector mosquito culex tritaeniorhynchus in australia using molecular diagnostics and morphology. *Parasites & vectors* **14**, 1–11.

[32] Babayan, S. A., Orton, R. J. & Streicker, D. G., 2018 Predicting reservoir hosts and arthropod vectors from evolutionary signatures in rna virus genomes. *Science* **362**, 577–580.

[33] Brierley, L. & Fowler, A., 2021 Predicting the animal hosts of coronaviruses from compositional biases of spike protein and whole genome sequences through machine learning. *PLoS Pathogens* **17**, e1009149.

[34] Mollentze, N., Babayan, S. & Streicker, D., 2021 Identifying and prioritizing potential human-infecting viruses from their genome sequences. *bioRxiv* pp. 2020–11.

[35] Poisot, T., Ouellet, M.-A., Mollentze, N., Farrell, M. J., Becker, D. J., Albery, G. F., Gibb, R. J., Seifert, S. N. & Carlson, C. J., 2021 Imputing the mammalian virome with linear filtering and singular value decomposition. *arXiv preprint arXiv:2105.14973* .

[36] Wardeh, M., Blagrove, M. S., Sharkey, K. J. & Baylis, M., 2021 Divide-and-conquer: machine-learning integrates mammalian and viral traits with network features to predict virus-mammal associations. *Nature Communications* **12**, 1–15.

[37] Young, F., Rogers, S. & Robertson, D. L., 2020 Predicting host taxonomic information from viral genomes: A comparison of feature representations. *PLoS computational biology* **16**, e1007894.

[38] Swei, A., Couper, L. I., Coffey, L. L., Kapan, D. & Bennett, S., 2020 Patterns, drivers, and challenges of vector-borne disease emergence. *Vector-Borne and Zoonotic Diseases* **20**, 159–170.

[39] dos Passos Cunha, M., Duarte-Neto, A. N., Pour, S. Z., Ortiz-Baez, A. S., Černỳ, J., de Souza Pereira, B. B., Braconi, C. T., Ho, Y.-L., Perondi, B., Sztajnbok, J. *et al.*, 2019 Origin of the são paulo yellow fever epidemic of 2017–2018 revealed through molecular epidemiological analysis of fatal cases. *Scientific reports* **9**, 1–10.

[40] Ryan, S. J., Carlson, C. J., Mordecai, E. A. & Johnson, L. R., 2019 Global expansion and redistribution of aedes-borne virus transmission risk with climate change. *PLoS neglected tropical diseases* **13**, e0007213.

[41] Ryan, S. J., Carlson, C. J., Tesla, B., Bonds, M. H., Ngonghala, C. N., Mordecai, E. A., Johnson, L. R. & Murdock, C. C., 2021 Warming temperatures could expose more than 1.3 billion new people to zika virus risk by 2050. *Global Change Biology* **27**, 84–93.

[42] Kraemer, M. U., Reiner, R. C., Brady, O. J., Messina, J. P., Gilbert, M., Pigott, D. M., Yi, D., Johnson, K., Earl, L., Marczak, L. B. *et al.*, 2019 Past

and future spread of the arbovirus vectors aedes aegypti and aedes albopictus. *Nature microbiology* **4**, 854–863.

[43] Messina, J. P., Brady, O. J., Golding, N., Kraemer, M. U., Wint, G. W., Ray, S. E., Pigott, D. M., Shearer, F. M., Johnson, K., Earl, L. *et al.*, 2019 The current and future global distribution and population at risk of dengue. *Nature microbiology* **4**, 1508–1515.

Figures

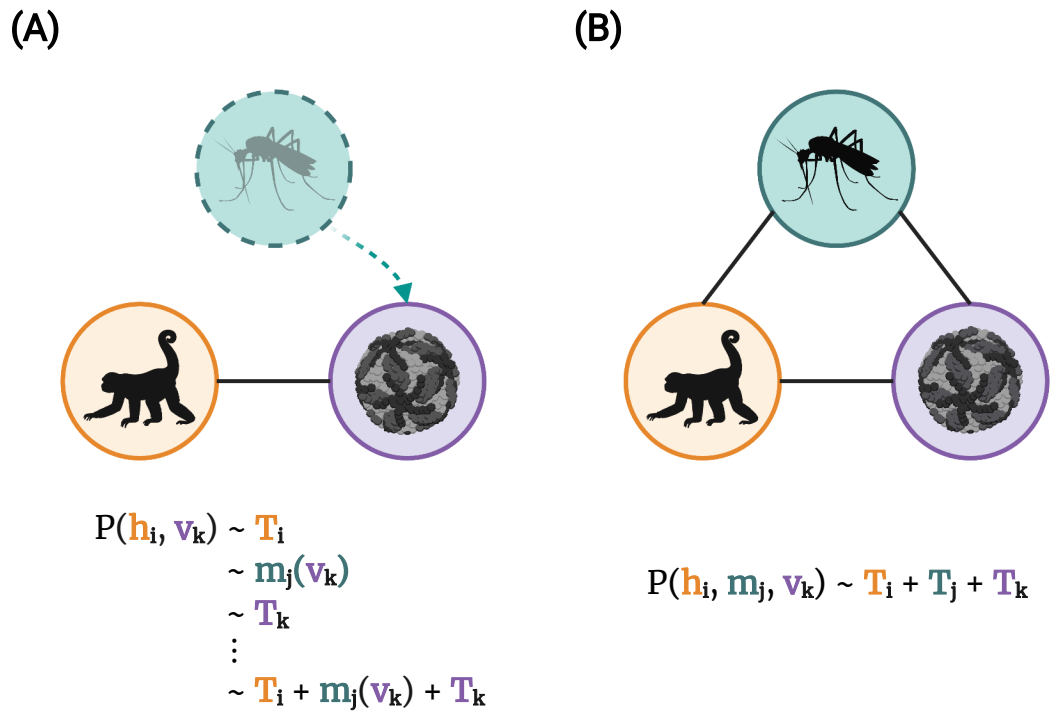

Figure 1: (A) Predicting host-virus associations (a bipartite network) based on host traits (T_h), virus traits (T_v), and vector communities ($m(v)$) associated with viruses, is a different problem than (B) predicting host-vector-virus associations (a tripartite network) based on host traits, vector traits, and virus traits. In this paper, we consider both solutions as approaches to end goals like forecasting potential novel associations or spillover scenarios

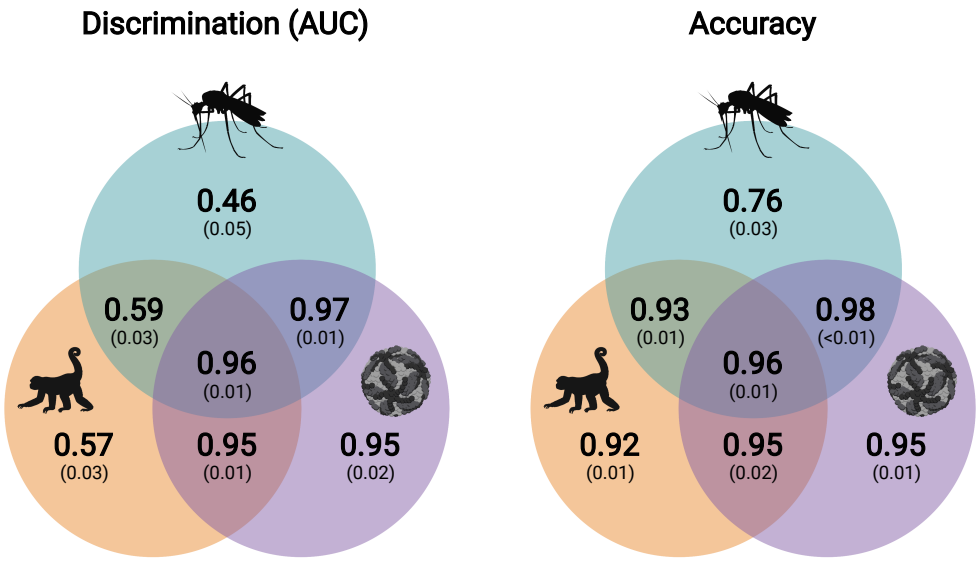

Figure 2: Model performance – quantified using AUC (left panel) and accuracy (right panel) – was highest when host, vector, and virus traits were included in the model (reported values are mean and standard deviation based on 20 model runs, assessing performance on a random 20% subset of the data). However, host-virus association model performance was not appreciably increased by the addition of 32 host trait covariates, suggesting that host-virus associations may be best predicted by considering information on the vector and the virus.

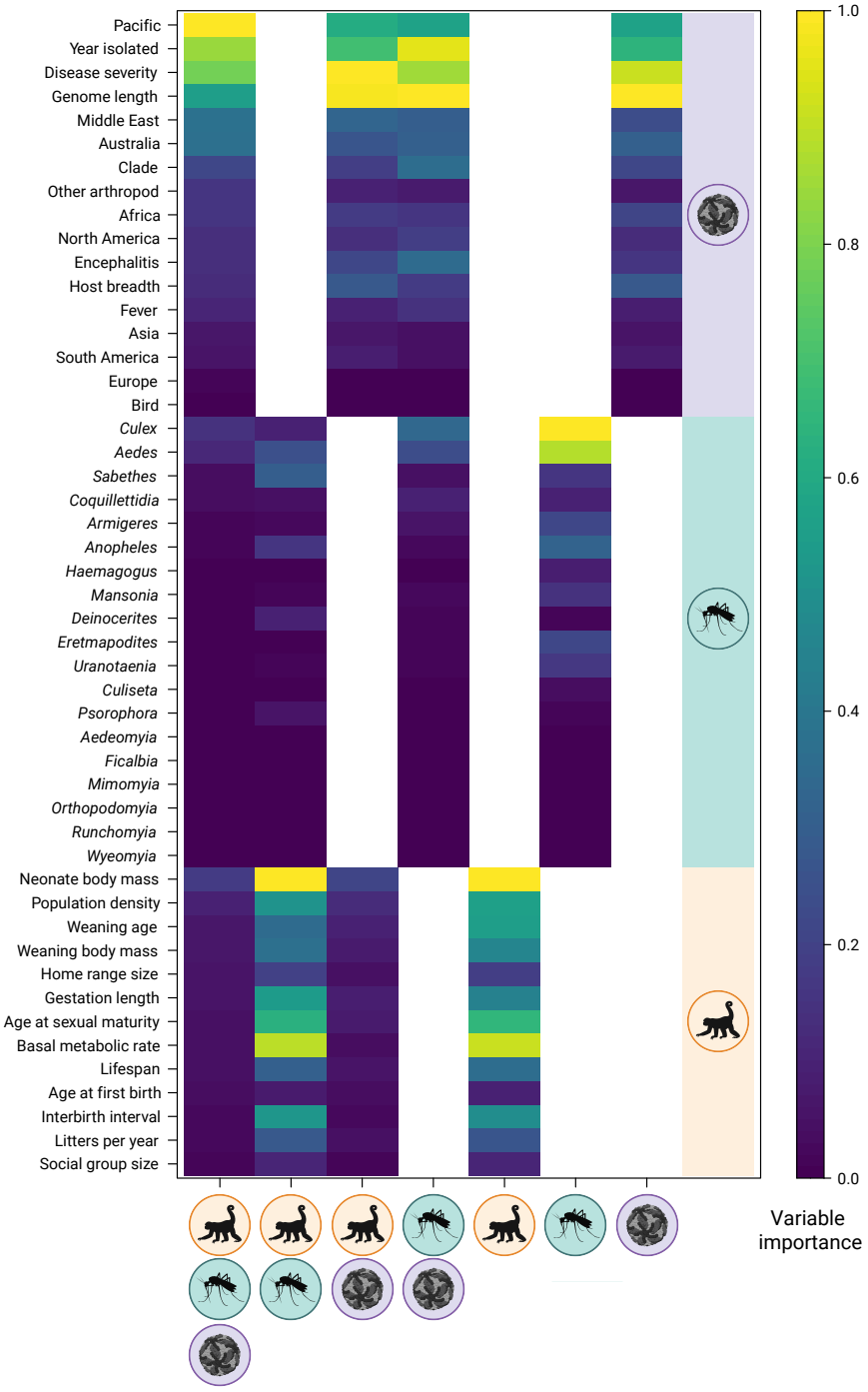

Figure 3: The relative importance of host (orange), vector (blue), and virus (purple) traits on predictive model performance. Each column corresponds to a different combination of these three trait groups, with the first column corresponding to the full model (as indicated at the bottom of each column using the glyphs). Variables are ordered based on the full model.

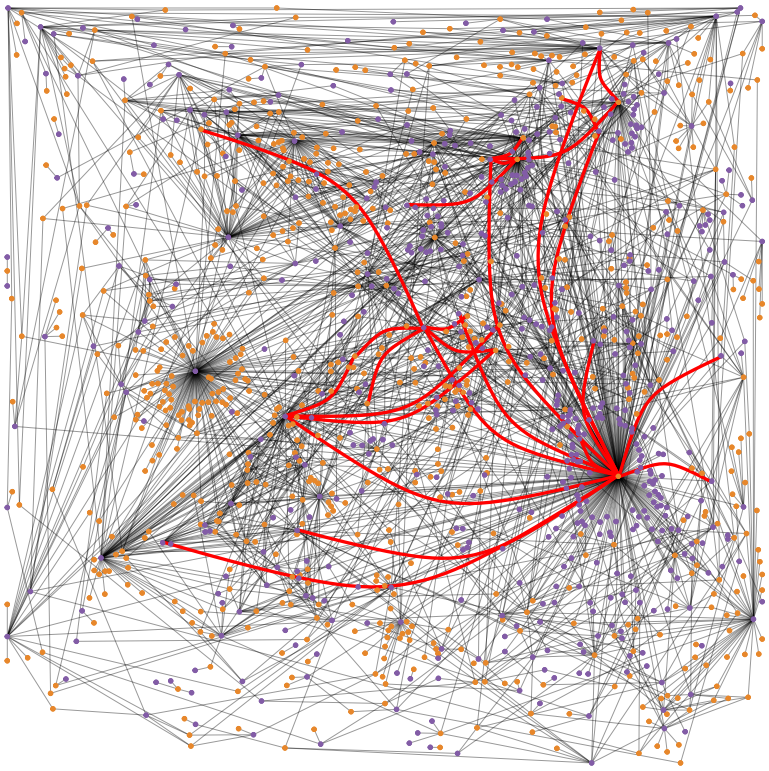

Figure 4: Full graph of host-virus associations (host species are in orange and viruses in purple), where links between host and virus species represent known associations. Red links are those which the full host-vector-virus cycle is known.

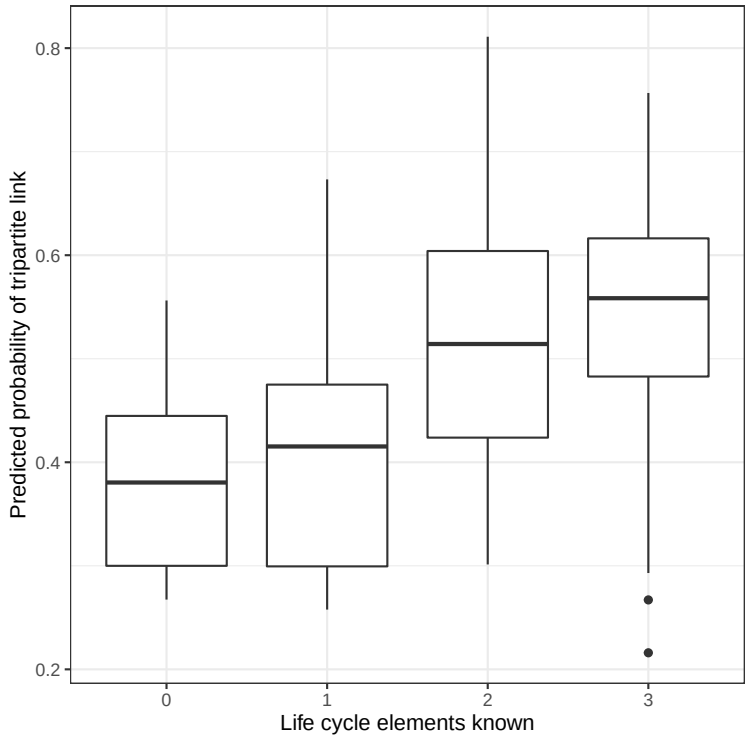

Figure 5: The tripartite model predicts a higher average probability for associations that have one or two links known (which are still not recorded as positive values in the training data) than those with no elements known to be possible. This suggests that the model is capable of more than just recapitulating the data, and is able to distinguish different levels of biological plausibility within unknown tripartite elements.

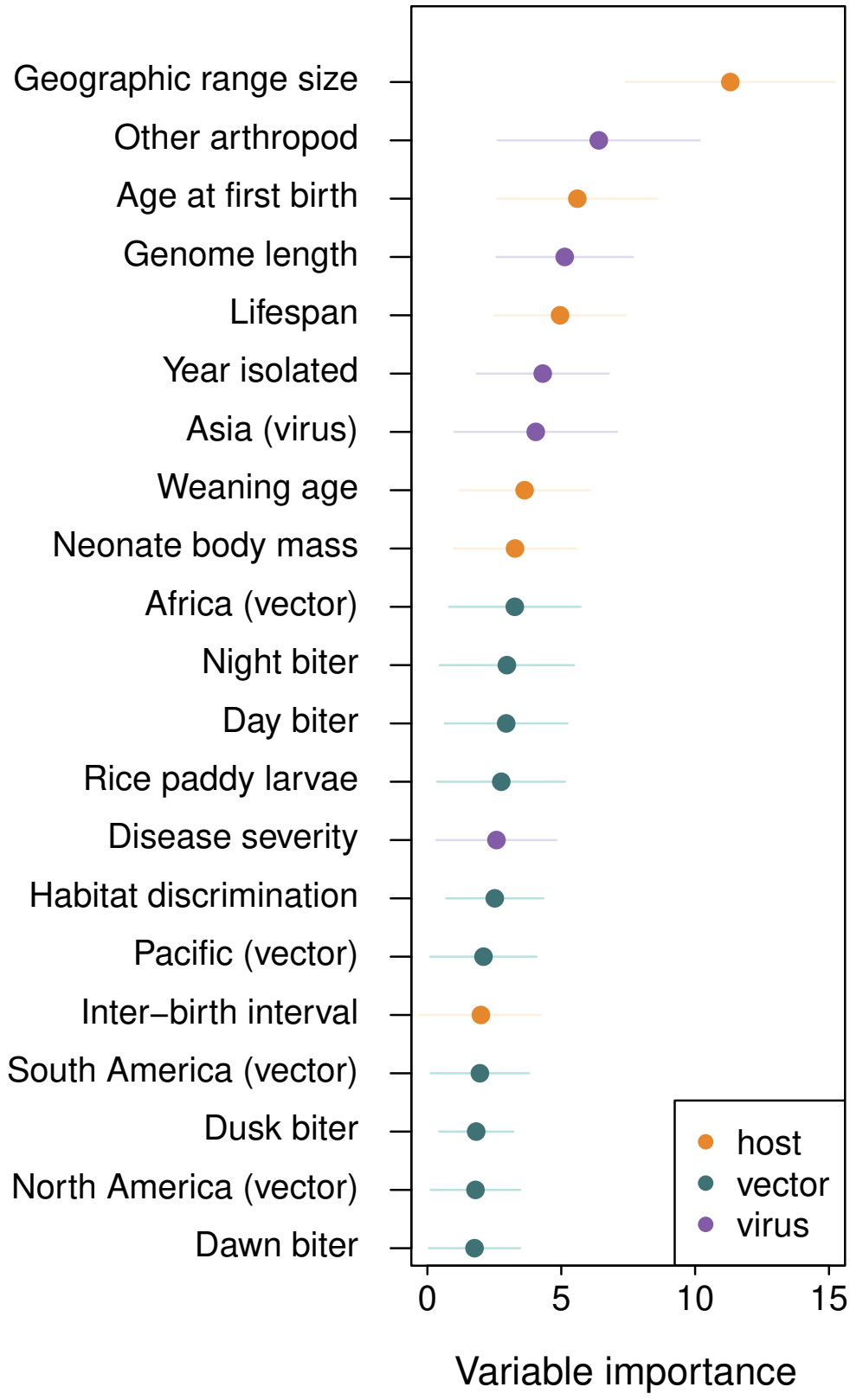

Figure 6: The relative importance of host (orange), vector (grey), and virus (blue) traits on predictive model performance in the tripartite model.

Tables

Table 1: **Top predicted epidemic cycles in humans.** All vectors are known to be human-biting; all viruses are known to be zoonotic based on either clinical or serological data.

Host	*Mosquito*	Virus	Prob
H. sapiens	*Culex pipiens*	Wesselsbron virus	0.81
H. sapiens	*Aedes aegypti*	West Nile virus	0.74
H. sapiens	*Aedes aegypti*	Japanese encephalitis virus	0.73
H. sapiens	*Culex pipiens*	Murray Valley encephalitis virus	0.73
H. sapiens	*Culex sitiens*	West Nile virus	0.72
H. sapiens	*Aedes scapularis*	West Nile virus	0.68
H. sapiens	*Mansonia uniformis*	West Nile virus	0.68
H. sapiens	*Anopheles coustani*	Wesselsbron virus	0.67
H. sapiens	*Culex pipiens*	Yellow fever virus	0.67
H. sapiens	*Aedes aegypti*	Ilheus virus	0.66
H. sapiens	*Aedes albopictus*	Ilheus virus	0.65
H. sapiens	*Anopheles hyrcanus*	Kokobera virus	0.62
H. sapiens	*Culex nigripalpus*	Wesselsbron virus	0.61
H. sapiens	*Aedes cantans*	Wesselsbron virus	0.61
H. sapiens	*Mansonia africana*	West Nile virus	0.61
H. sapiens	*Culex perexiguus*	Wesselsbron virus	0.60
H. sapiens	*Culex thalassius*	Wesselsbron virus	0.60
H. sapiens	*Culex gelidus*	West Nile virus	0.60
H. sapiens	*Culex annulirostris*	St. Louis encephalitis virus	0.59
H. sapiens	*Anopheles pharoensis*	West Nile virus	0.59

Table 2: **Top predicted enzootic cycles.** All mammals in the top 20 are either species found alongside humans (cows, sheep, pigs, and rats) or easily-sampled species from eastern North America (deer, raccoons, and possums).

Host	*Mosquito*	Virus	Prob
Sus scrofa	*Aedes albopictus*	West Nile virus	0.70
Sus scrofa	*Mansonia uniformis*	Wesselsbron virus	0.69
Didelphis virginiana	*Aedes aegypti*	Wesselsbron virus	0.67
Didelphis virginiana	*Aedes albopictus*	Wesselsbron virus	0.67
Sus scrofa	*Anopheles coustani*	Wesselsbron virus	0.66
Sus scrofa	*Culex quinquefasciatus*	Japanese encephalitis virus	0.65
Procyon lotor	*Culex tritaeniorhynchus*	West Nile virus	0.62
Odocoileus virginianus	*Anopheles pharoensis*	Wesselsbron virus	0.62
Procyon lotor	*Culex pipiens*	Japanese encephalitis virus	0.61
Didelphis virginiana	*Aedes aegypti*	West Nile virus	0.59
Odocoileus virginianus	*Aedes albopictus*	St. Louis encephalitis virus	0.56
Bos taurus	*Aedes vexans*	West Nile virus	0.56
Bos taurus	*Culex tritaeniorhynchus*	West Nile virus	0.55
Procyon lotor	*Culex annulirostris*	West Nile virus	0.55
Macropus agilis	*Culex tritaeniorhynchus*	Murray Valley encephalitis virus	0.54
Bos taurus	*Anopheles maculipennis*	Wesselsbron virus	0.54
Ovis aries	*Culex quinquefasciatus*	Ilheus virus	0.53
Procyon lotor	*Culex tarsalis*	West Nile virus	0.52
Rattus rattus	*Aedes aegypti*	Zika virus	0.51
Odocoileus virginianus	*Culex tritaeniorhynchus*	Banzi virus	0.51

Supplemental Material

Predicting the tripartite network of mosquito-borne disease

Trait data

Trait data were compiled from a variety of sources, with host trait data coming from PanTHERIA [21], and vector and virus trait data from Evans et al. 2017 [13].

Table S1: Host, vector, and virus covariates considered in the models of host-virus (`h-v` column) and host-vector-virus (`h-m-v` column) associations. See the Pantheria documentation (https://esapubs.org/archive/ecol/E090/184/metadata.htm) for more information on host trait variables.

Taxa	Variable	Units	Definition	h-v	h-m-v
Host					
	Lifespan	days	Maximum observed lifespan	☑	☑
	Age at sexual maturity	days	Age at which individual is sexually mature	☑	☑
	Home range size	km^2	Area used by individual for daily tasks on average	☑	
	Gestation length	days	Period of time young are gestated	☑	☑
	Neonate body mass	grams	Average neonate body mass	☑	☑
	Population density	n / km^2	Number of individuals per unit area, on average	☑	
	Age at first birth	days	Age at which females give birth to their first litter	☑	☑
	Litters per year	n / year	Average number of litters per year	☑	
	Max lifespan	months	Longest observed lifespan	☑	☑
	Basal metabolic rate	mLO_2 / hr	Individual metabolic rate	☑	

	Interbirth interval	months	Period in between reproductive bouts	☑	☑
	Age at eye opening	days	Time when neonates open eyes	☑	
	Social group size	count	Number of individuals per social group	☑	
	Adult forearm length	mm	Length of adult forearm	☑	
	Dispersal age	days	Age at which young leave parents	☑	
	Neonate head-body length	mm	Body length of neonates	☑	
	Weaning age	days	Period of time when young stop weaning	☑	☑
	Weaning body mass	grams	Mass of young during weaning	☑	☑
Vector					
	Mosquito genus	numeric	Number of mosquito species of genus that vector a given virus		☑
	Human biter	1/0	Vector bites humans	☑	
	Host breadth	count	Number of host species bitten	☑	
	Non-primate mammals	1/0	Are non-primate mammals bitten	☑	
	Geographic range	count	Number of countries species collected	☑	
	Geographic location	-	Could include any or all of the following; Africa, Middle East, Australia, Pacific, Asia, Europe, North America, South America	☑	
	Biting behavior	-	Timing of biting behavior. Can be; dawn, day, dusk, and/or night	☑	
	Artificial container	1/0	Vector breeds in artificial containers	☑	

	Oviposition site	-	Larval site. Could include one or all of; treehole, container, pond, rockhole, marsh, swamp, ground pool, or rice paddy	☑	
	Permanent habitat	1/0	Species uses permanent habitat	☑	
	Habitat discrimination	count	number of habitat types	☑	
	Urban preference	1/0	vector shows urban preference	☑	
	Indoor preference	1/0	vector shows indoor preference	☑	
	Viral range	count	Number species within genus to harbor virus	☑	
Virus					
	Average genome length	numeric	Length of viral genome	☑	☑
	Geographic location	-	Could include any or all of the following; Africa, Middle East, Australia, Pacific, Asia, Europe, North America, South America	☑	☑
	Clade	–	Viral clade (roman numerals)	☑	
	Year isolated	year	Virus isolation year	☑	☑
	Other arthropod	1/0	Vectored by other arthropods	☑	
	Host breadth	count	number of known hosts	☑	☑
	Encephalitis	1/0	Virus causes encephalitis	☑	☑
	Fever	1/0	Virus causes fever	☑	☑
	Disease severity	numeric	How severe is disease	☑	☑
	Bird host	1/0	Virus infects birds		☑

What if we consider all vector-borne viruses?

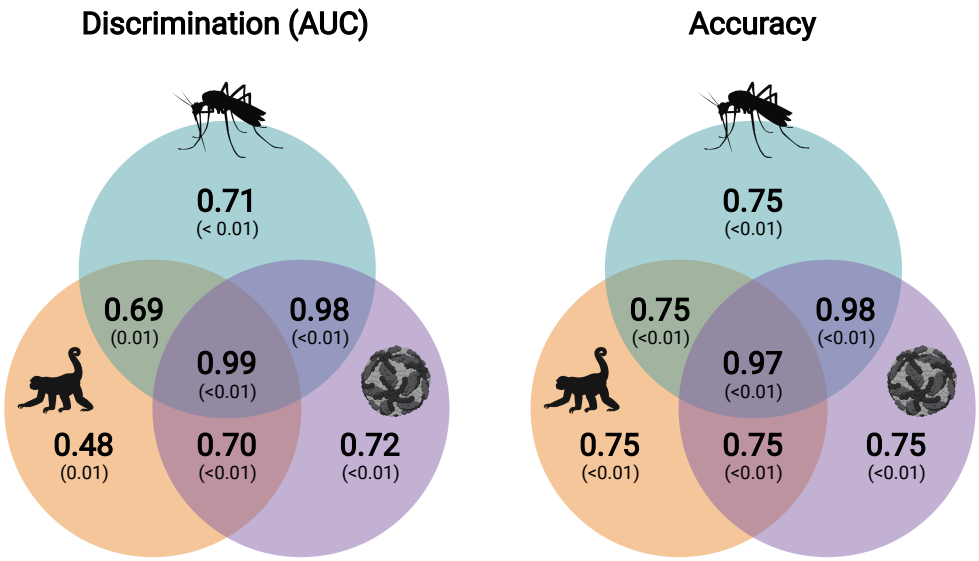

Figure S1: Model performance – quantified using AUC (left panel) and accuracy (right panel) – was highest when host, vector, and virus traits were included in the model (reported values are mean and standard deviation based on 20 model runs, assessing performance on a random 20% subset of the data). However, host-virus association model performance was not appreciably increased by the addition of vector data compared to just host and vector traits (AUC = 0.98).

Different models and similar predictions

When predicting host-virus associations, the different models had quite different variable importance values, apart from obviously having different explanatory variables. One question we had was whether models trained on different covariates would not only have similar overall performance, but identify the same likely host-virus associations as other models. To explore this graphically, we generated a correlation matrix (Figure S2), where we find strong positive relationships between different model predictions. Interestingly, the least positive correlation was from the full model, suggesting that the predictions from the full model differed from models which consisted of nested subsets of the same features as in the full model.

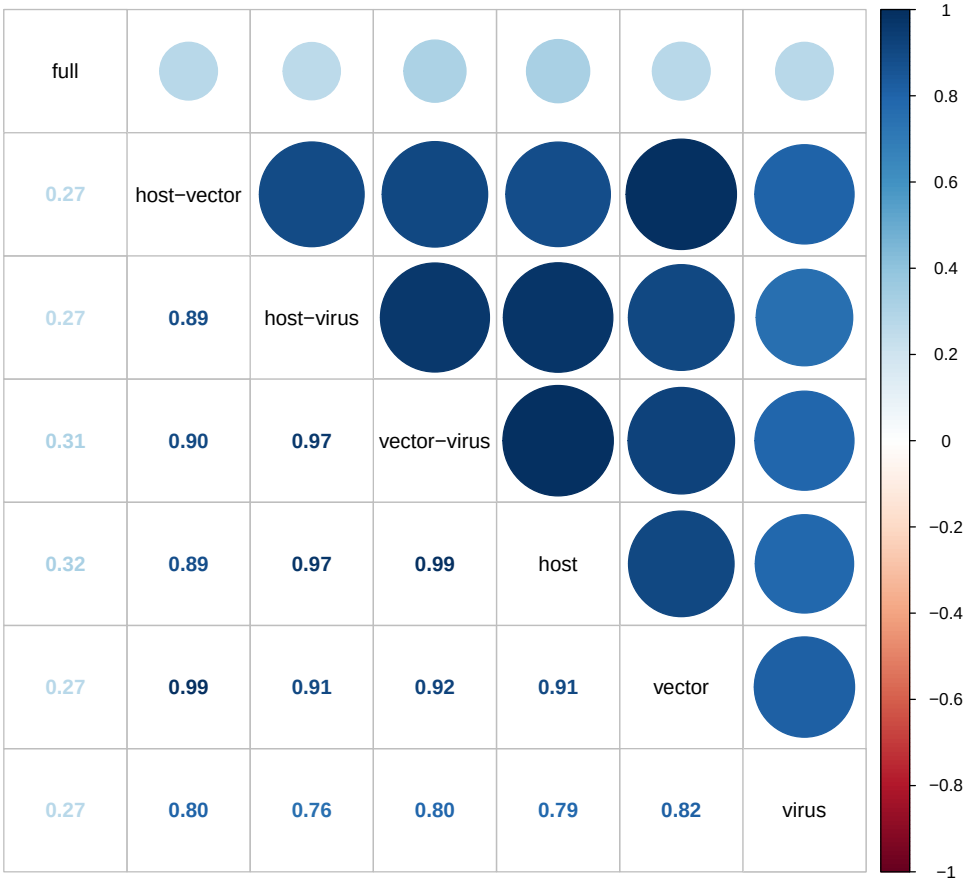

Figure S2: Correlation matrix between model predictions of the full set of subset models including different combinations of host, vector, and virus traits. The full model, including all traits, resulted in the predictions that were most weakly related to the other model predictions, though this model had similar performance as other models (see main text Figure 2). Lower triangle values and color scale correspond to Pearson's correlation coefficient values.