




RESEARCH ARTICLE

Predicting missing links in global host–parasite networks

Maxwell J. Farrell^{1,2,3}  | Mohamad Elmasri⁴ | David A. Stephens⁴  |
T. Jonathan Davies^{5,6} 

¹Department of Biology, McGill University, QC, Canada; ²Ecology & Evolutionary Biology Department, University of Toronto, Toronto, ON, Canada; ³Center for the Ecology of Infectious Diseases, University of Georgia, Athens, GA, USA; ⁴Department of Mathematics & Statistics, McGill University, QC, Canada; ⁵Botany, Forest & Conservation Sciences, University of British Columbia, Vancouver, BC, Canada and ⁶African Centre for DNA Barcoding, University of Johannesburg, Johannesburg, South Africa

Correspondence

Maxwell J. Farrell

Email: maxwell.farrell@utoronto.ca

Funding information

McGill University

Handling Editor: Matthew Barbour**Abstract**

1. Parasites that infect multiple species cause major health burdens globally, but for many, the full suite of susceptible hosts is unknown. Predicting undocumented host–parasite associations will help expand knowledge of parasite host specificities, promote the development of theory in disease ecology and evolution, and support surveillance of multi-host infectious diseases. The analysis of global species interaction networks allows for leveraging of information across taxa, but link prediction at this scale is often limited by extreme network sparsity and lack of comparable trait data across species.
2. Here we use recently developed methods to predict missing links in global mammal–parasite networks using readily available data: network properties and evolutionary relationships among hosts. We demonstrate how these link predictions can efficiently guide the collection of species interaction data and increase the completeness of global species interaction networks.
3. We amalgamate a global mammal host–parasite interaction network (>29,000 interactions) and apply a hierarchical Bayesian approach for link prediction that leverages information on network structure and scaled phylogenetic distances among hosts. We use these predictions to guide targeted literature searches of the most likely yet undocumented interactions, and identify empirical evidence supporting many of the top ‘missing’ links.
4. We find that link prediction in global host–parasite networks can successfully predict parasites of humans, domesticated animals and endangered wildlife, representing a combination of published interactions missing from existing global databases, and potential but currently undocumented associations.
5. Our study provides further insight into the use of phylogenies for predicting host–parasite interactions, and highlights the utility of iterated prediction and targeted search to efficiently guide the collection of information on host–parasite interactions. These data are critical for understanding the evolution of host specificity, and may be used to support disease surveillance through a process of predicting missing links, and targeting research towards the most likely undocumented interactions.

KEYWORDS

disease ecology, host–parasite interactions, infectious diseases, macroecology, phylogenetics

1 | INTRODUCTION

Most disease-causing organisms of humans and domesticated animals can infect multiple host species (Cleaveland et al., 2001; Taylor et al., 2001). This has ramifications for biodiversity conservation (Farrell et al., 2021; Smith et al., 2009) and human health via direct infection, food insecurity and diminished livelihoods (Grace et al., 2012). Despite the severe burdens multi-host parasites can impose, we do not know the full range of host species for the majority of infectious organisms (Dallas et al., 2017). Currently, parasite host ranges are best described by host–parasite interaction databases that are largely compiled from primary academic literature. These databases gain strength by collating data across a large diversity of hosts and parasites, and have been used to identify macroecological patterns of infectious diseases (Stephens et al., 2016, 2017), define life histories influencing host specificity (Park, 2019; Park et al., 2018) and predict the potential for zoonotic spillover (Olival et al., 2017). However, global databases are known to be incomplete, with some estimated to be missing up to ~40% of host–parasite interactions among the species sampled (Dallas et al., 2017). While even incomplete data can offer important insights into the structure of host–parasite interaction networks, our ability to make accurate predictions decreases when data are missing.

Filling in knowledge gaps, and building more comprehensive global databases of host–parasite interactions enhance our insights into the ecological and evolutionary forces shaping parasite biodiversity. These insights include the role of network structure for transmission (Gomez et al., 2013; Pulosof et al., 2015), the nature of highly implausible host–parasite interactions (Morales-Castilla et al., 2015), the drivers of parasite richness (Ezenwa et al., 2006; Huang et al., 2015; Kamiya et al., 2014; Nunn et al., 2003) and parasite sharing across hosts (Albery et al., 2020; Braga et al., 2015; Davies & Pedersen, 2008; Huang et al., 2014; Luis et al., 2015), the association between host specificity and virulence (Farrell & Davies, 2019; Shwab et al., 2018) and global estimates of parasite diversity (Carlson et al., 2019). While determining the outcome of a given interaction ultimately requires moving beyond the binary associations provided by current databases, identifying documented and likely host–parasite associations is a critical first step.

Recent efforts have used species traits to predict reservoirs of zoonotic diseases (Han et al., 2015) and identify wildlife hosts of globally important viruses (Han et al., 2016; Pandit et al., 2018). Studies of global host–virus interactions have also been used to predict the structure of viral sharing networks (Albery et al., 2020; Wardeh et al., 2020). Sharing networks model interactions as unipartite networks in which hosts are connected by having at least one parasite in common. These may be derived from bipartite interaction data, but the derived networks effectively consider parasites as interchangeable. While sharing networks can identify host profiles for particular parasite groups or host species that promote parasite sharing across hosts, the ability to predict individual host–parasite interactions is limited (Becker et al., 2022). An alternative is to treat hosts and parasites as two separate interacting classes that can be

represented in a bipartite network (Albery et al., 2021). Bipartite network models can predict new links based solely on the structure of the observed network, or incorporate node-level covariates, such as species traits (Becker et al., 2022; Dallas et al., 2017).

Algorithms such as recommender systems can identify probable links in a variety of large networks (Ricci et al., 2011). These models attempt to capture two key properties of real-world networks: the scale-free behaviour of interactions (shown by the degree distributions in Figure SM 1) and local clustering of interactions (Watts & Strogatz, 1998). However, they also tend to predict that nodes with many documented interactions are more likely to associate with other highly connected nodes in the network (a behaviour described as ‘the rich-get-richer’). In the context of host–parasite interactions, the number of interactions per species will vary due to ecological or evolutionary processes, and may also be influenced by research effort. This is commonly seen in the studies of parasite species richness in which sampling effort explains a large portion of the variation across hosts (Ezenwa et al., 2006; Kamiya et al., 2014; Lindenfors et al., 2007; Nunn et al., 2003; Olival et al., 2017). Models that directly or indirectly use the number of documented interactions per species to make predictions cannot easily adjust for these sampling biases. Thus, while these affinity-based models may be highly tractable for large networks, they are also sensitive to uneven sampling across nodes, and may re-enforce observation biases.

Predictions from trait-informed network models are less sensitive to missingness, but require data on ecological and functional traits of interacting species (Bartomeus et al., 2016; Dallas et al., 2017; Gravel et al., 2013). Thus, while these approaches work well for smaller networks (Dallas et al., 2017), they can scale poorly to global-scale ecological datasets in which comparable traits are unavailable for all species (Morales-Castilla et al., 2015). In the absence of comprehensive trait data, we can incorporate evolutionary relationships among hosts to add biological realism to network-based predictions. Phylogenetic trees represent species' evolutionary histories, and branch lengths of these trees provide a measure of expected similarities among species (Wiens et al., 2010). Hosts may be associated with a parasite through inheritance from a common ancestor, or as a result of host shifts (Page, 1993). In both processes we expect closely related species will host similar parasite assemblages (Davies & Pedersen, 2008).

Here we apply a link prediction model for bipartite ecological networks that combines properties of affinity-based models with information on host phylogeny (Elmasri et al., 2020), to a massive global-scale host–parasite network for mammals. Incorporating host phylogeny into ‘rich-get-richer’ (i.e. scale-free) interaction models adds biological structure, allowing for more realistic predictions with otherwise limited covariate data. This approach is particularly well-suited to link prediction in large global host–parasite networks as it does not require trait data and allows accurate predictions of missing links in extremely sparse networks using only the structure of the observed host–parasite associations and scaled evolutionary relationships among hosts. We demonstrate how model predictions can efficiently guide efforts to fill in missing links, show their

geographical distributions and update existing global networks using historical and recently published interactions.

2 | MATERIALS AND METHODS

2.1 | Data

To generate a network of documented host-parasite interactions for mammals, we amalgamated four major global host-parasite databases (Gibson et al., 2005; Olival et al., 2017; Stephens et al., 2017; Wardeh et al., 2015). These are derived from primary literature, genetic sequence databases and natural history collections, and report host and parasite names as Latin binomials. The Global Mammal Parasite Database 2.0 (GMPD) (Stephens et al., 2016) contains records of disease-causing organisms (viruses, bacteria, protozoa, helminths, arthropods and fungi) in wild ungulates (artiodactyls and perrisodactyls), carnivores and primates drawn from over 2,700 literature sources published through 2010 for ungulates and carnivores, and 2015 for primates. The static version of the Enhanced Infectious Disease Database (EID2) (Wardeh et al., 2015) contains 22,515 host-pathogen interactions from multiple kingdoms based on evidence published between 1950 and 2012 extracted from the NCBI Taxonomy database, NCBI Nucleotide database and PubMed citation and index. Due to the semi-automated procedure used to generate this database, some commensal or mutualistic interactions are included. The database does not contain metadata to filter these interactions, but they are assumed to be rare relative to parasitic interactions (Wardeh et al., 2015). The Host-Parasite Database of the Natural History Museum, London (Gibson et al., 2005) contains over a quarter of a million host-parasite records for helminth parasites extracted from 28,000 references published after 1922, and is digitally accessible via the R package `HELMINTHR` (Dallas, 2016). Finally, Olival et al. (2017) compiled a database of 2,805 mammal-virus associations for every recognized virus found in mammals, representing 586 unique viral species and hosts from 15 mammalian orders. These source databases were then combined into a single database, harmonizing taxonomy of hosts and parasites (full details in SM 1.1). We treat interactions as binary (0/1) for a given host-parasite pair as the sources do not explicitly indicate the role a host species plays in parasite transmission, but instead approximate host exposure and susceptibility to infection.

The resulting network includes 29,112 documented associations among 1,835 host and 9,149 parasite species (Figure 1). To our knowledge this constitutes the largest mammal host-parasite interaction network currently available, and includes parasites from diverse groups including viruses, bacteria, protozoa, helminths, arthropods and fungi, in wild, domestic and human hosts. The resulting matrix is quite sparse, with ~0.17% of the ~16.8 million possible links having documented interactions. Humans are documented to associate with 2,064 parasites (47% of which associate with another mammal in the database), and comprise 7% of all interactions. Parasite species are largely represented by helminths (63.9%), followed by bacteria (13.1%) and viruses (7.89%). The degree distribution (number of documented

interactions per species) varies considerably and is shown to be linear on the log scale for both hosts and parasites (Figure SM 1).

2.2 | Statistical analyses

We apply the bipartite link prediction model of Elmasri et al. (2020) to the amalgamated dataset. The model has three variants: the 'affinity' model which generates predictions based only on the number of observed interactions for each host and parasite, the 'phylogeny' model which is informed only by host evolutionary relationships (here taken from Fritz et al. (2009)) and the 'combined' model which layers both components (termed 'full' model in Elmasri et al. (2020)). The affinity model is fit by preferential attachment whereby hosts and parasites that have many interacting species in the network are assigned higher probabilities of forming novel interactions. The phylogeny model uses the similarity of host species based on evolutionary distances to assign higher probability to parasites interacting with hosts closely related to their documented host species, and lower probability of interacting with hosts that are distantly related. To account for uncertainty in the phylogeny, and allow the model to place more or less emphasis on recent versus deeper evolutionary relationships, we fit a tree scaling parameter (η) based on an accelerating-decelerating model of evolution (Harmon et al., 2010). This transformation allows for changes in the relative evolutionary distances among hosts and was shown to have good statistical properties for link prediction in a subset of the GMPD (Elmasri et al., 2020). We apply these three models to the full dataset. The tree scaling parameter is applied across the whole phylogeny, but since the importance of recent versus deep evolutionary relationships among hosts is likely to vary across parasite types (Park et al., 2018), we additionally run the models on the dataset subset by parasite taxonomy (arthropods, bacteria, fungi, helminths, protozoa and viruses). For all models, we used 10-fold cross-validation to prevent over-fitting during parameter estimation, and to assess model performance when predicting links internal to the dataset (see SM 1.2 for details).

2.3 | Targeted literature searches

We identified the top 10 most likely links in each model-data subset combination which were not documented in the original data. These were used to guide searches of primary and grey literature for evidence of associations. Searches were conducted in Google Scholar by using both the host and parasite Latin binomials in quotes and separated by the AND Boolean operator (e.g. 'Gazella leptoceros' AND 'Nematodirus spathiger'). If this returned no hits, we searched using alternative names, and then used the standard Google engine to identify grey literature sources. For models run on the full dataset, we also investigated the top 10 links for domesticated mammals (as defined by Clutton-Brock (1999) and harmonized to Wilson and Reeder (2005): *Bison bison*, *Bos* sp., *Bubalus bubalis*, *Camelus* sp.,

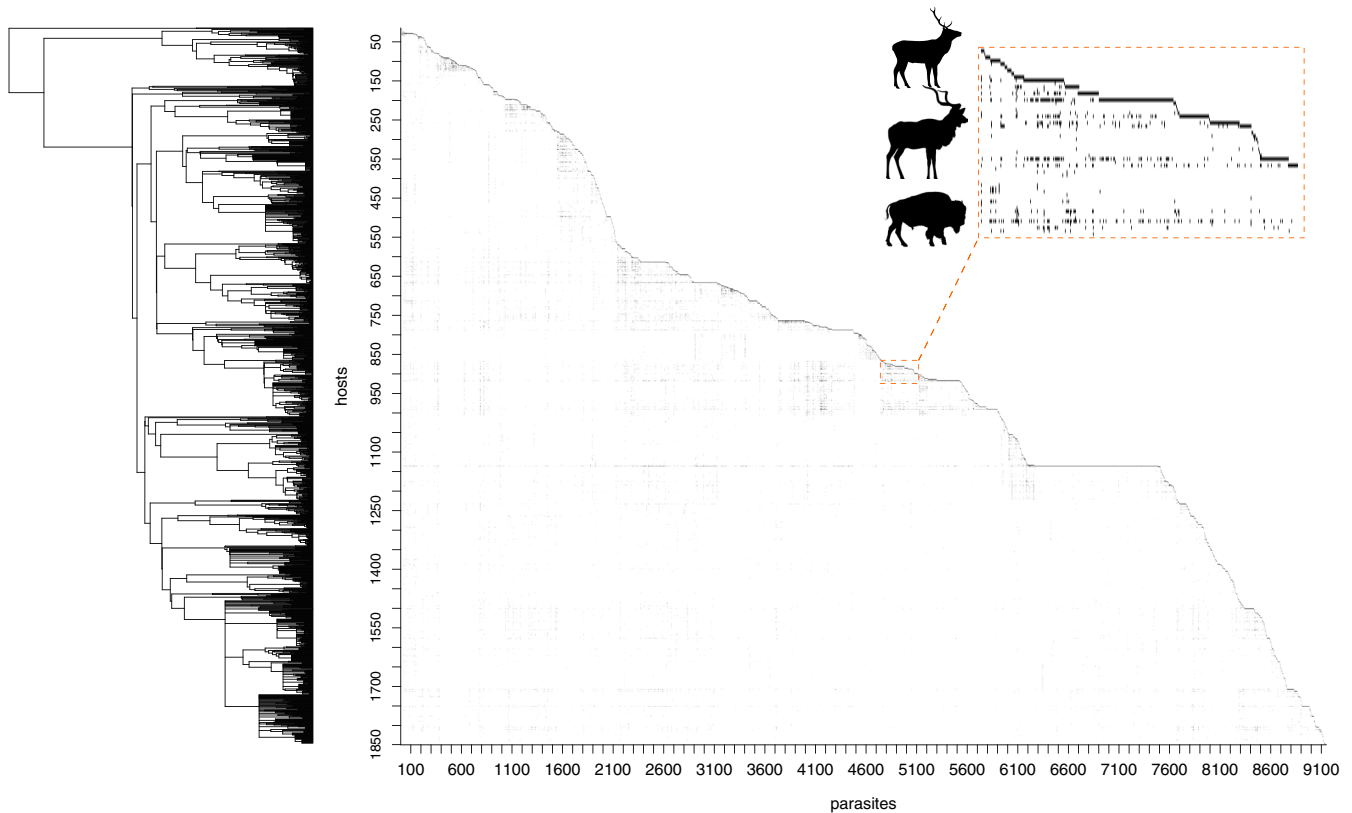


FIGURE 1 Phylogeny of host species (pruned from the Fritz et al. (2009) dated supertree for mammals), and host–parasite association matrix showing hosts ordered according to the phylogeny. Axes represent hosts (rows) and parasites (columns). The orange rectangles display an expanded subset of the matrix

Capra hircus, *Canis lupus*, *Cavia porcellus*, *Equus asinus*, *Equus caballus*, *Felis catus*, *Felis silvestris*, *Lama glama*, *Mus musculus*, *Oryctolagus cuniculus*, *Ovis aries*, *Rangifer tarandus*, *Rattus norvegicus*, *Rattus rattus*, *Sus scrofa*, *Vicugna vicugna*, and wild host species separately. For domesticated animals and humans, if the Latin binomials returned no hits, the search strategy was repeated using host common names (e.g. ‘human’, ‘pig’, ‘horse’).

We considered any physical, genetic or serological identification of a parasite infecting a given host species as evidence of an association. Exceptions included (a) situations where parasite identification was stated as uncertain due to known serological cross-reactivity, (b) absence of clear genetic similarity to known reference sequences or (c) unconfirmed visual diagnosis made from afar.

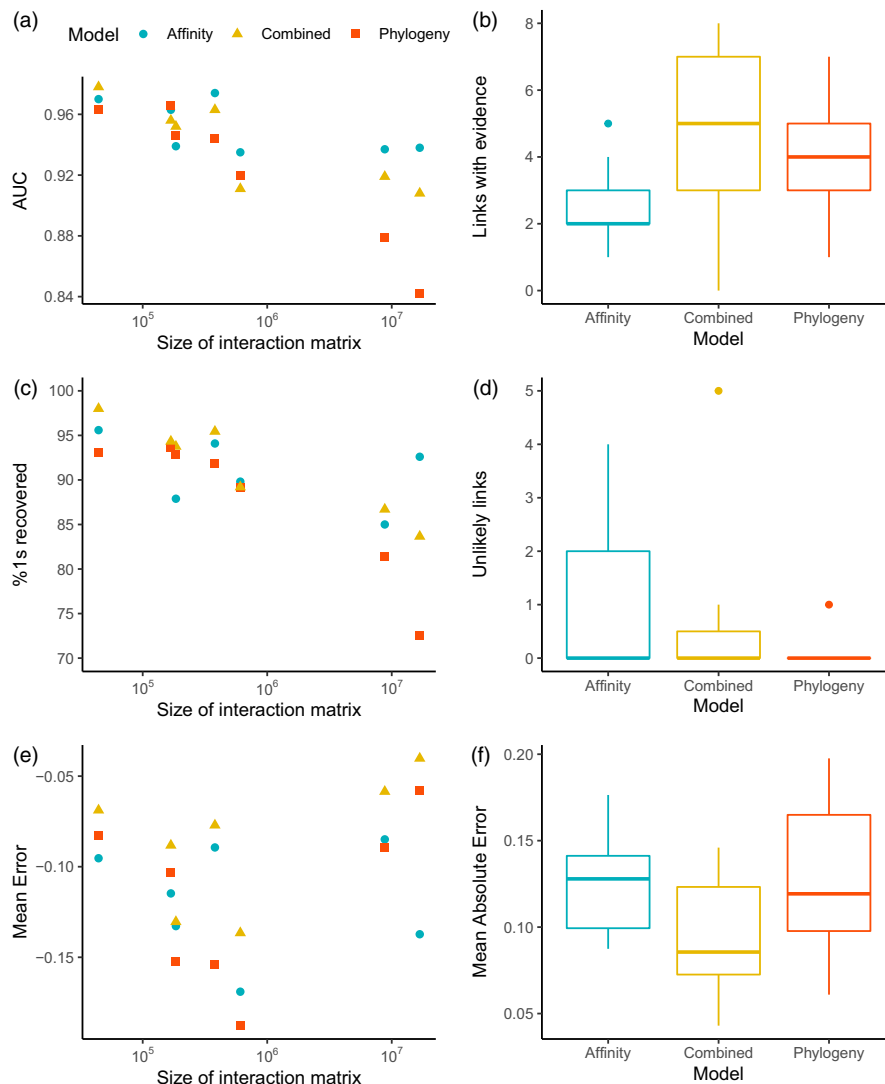
In total, we generated 27 sets of ‘top 10’ highly likely undocumented links to target, because of overlap in the top links among models run across data subsets, this reduced to 177 unique links investigated for published evidence of infection. Missing links were classified as unlikely when involving ecological mismatch between host and parasite, such as trophically transmitted parasites infecting the wrong trophic level or unlikely ingestion pathway ($n = 16$). As our goal is to predict potential susceptibility, lack of known current geographic overlap among wide ranging hosts and parasites was not considered sufficient to classify a link as unlikely.

3 | RESULTS

When assessing models using 10-fold cross-validation, all models performed well when predicting links internal to the data. Area under the receiver operating characteristic curve (AUC) values ranged from 0.84 to 0.98 (AUC of 1 signifies perfect predictive accuracy), and between 72.54% and 98.00% of the held-out documented interactions were successfully recovered (Figure 2, Table S1; see Figure SM 2 for posterior interaction matrices for the full dataset). Model performance as measured by these metrics tended to decrease with the size of the interaction matrix (Figure 2a,c). We also investigate predictive error as measured by mean error (ME; Figure 2e), mean absolute error (MAE; Figure 2f) and root mean square error (RMSE; Table S1). The negative values of ME for all models indicate that they are consistently ‘over’ predicting interactions relative to numbers of observed interactions in each dataset, as we would expect given assumed missing interactions in the data.

Our goal was to rank predicted probabilities of links to identify which missing links may be most likely, rather than thresholding the predictions to produce a snapshot of a ‘complete’ network. However, depending on context, users might wish to threshold predicted interaction probabilities to increase or decrease the prediction of new 1s. An alternative approach is to choose a threshold that maximizes precision and recall, or their harmonic mean (F1 score). The area

FIGURE 2 Model diagnostic plots, including internal predictive performance after 10-fold cross-validation (a, c, e, f) (see SM 1.2 for details), and the results of targeted literature searches for the top 10 undocumented links per model (b, d). Panel (a) shows area under the receiver operating characteristic curve (AUC), panel (c) shows the per cent documented interactions (1s) correctly recovered from the held-out portion and panel (e) shows the mean predicted error (bias), each in relation to the size of the interaction matrix. Panel (f) shows boxplots of mean absolute error by model type. Panels (b) and (d) show boxplots of the number of links with published evidence external to the original dataset (b), and the number of unlikely links based on ecological mismatch (d) out of the top 10 most probable yet undocumented interactions for each of the three model types (affinity, combined and phylogeny) run across the full dataset and each of the models' subset by parasite type (arthropods, bacteria, fungi, helminths, protozoa and viruses)



under the precision-recall curve and F1 score ranged widely across datasets and models (Table S2), but in many cases performed on par with state-of-the-art deep learning models applied to problems with similar imbalance (Johnson & Khoshgoftaar, 2019). The combined models consistently showed the smallest bias and highest accuracy when measured via MAE and RMSE (Table S1). Unlike AUC, ME did not decrease with size of the interaction matrix, instead achieving highest accuracy when using the combined model on the full dataset, indicating that including information across parasite types may help to reduce potential false positives compared to models restricted to one parasite type. While these measures act as alternative performance metrics to AUC and % 1s, model performance is still difficult to assess with respect to false positives without inclusion of 'true negative' (i.e. experimental infection) data or comprehensive literature review identifying highly unlikely interactions.

The models of Elmasri et al. (2020) generate relative interaction probabilities for an entire bipartite matrix. The intention of the approach is to rank undocumented interactions as a means for directing future study, as we have done here with our targeted literature reviews. However, if there is concern about potential false positives

in the top-ranked predictions, it may be possible to identify a cut-off below which additional study is less likely to reveal new interactions. One approach would be to use an external model such as that described in Dallas et al. (2017), which estimates either the fraction of undocumented interactions, or expected numbers of links per host and parasite. Alternatively, the top-ranked predictions can be binned and the fraction of interactions already observed can be plotted to visualize potential drop-offs in yield as less likely interactions are investigated. We conduct this as a post hoc analysis (SM 3) and find that all models tend to have the highest recovery of observed interactions among top-ranked interactions. However, the combined model displays the most consistent drop-off in proportion of observed interactions as a function of prediction rank, which starts to level off around 5,000 interactions. This indicates that the predicted interaction probabilities are able to segregate between observed and unobserved interactions, offering an approach to limit potential false positives for studies that wish to predict the structure of the entire interaction network rather than just rank top undocumented links.

As would be expected, the affinity-only model tended to predict links between species with many previously documented

associations, but this behaviour was decreased in the combined model, and largely absent in the phylogeny-only model (Figure 3, SM 2.2). To further explore how each of the link prediction models propagated potential research biases, we compared the predicted probabilities per interaction to the degree product (calculated per host–parasite interaction by multiplying the observed host degree by the observed parasite degree). The degree product represents the basic expectation of a given interaction based on host and parasite affinities. Comparing the observed degree products to the predicted interaction probabilities, we can see that the predictions from the affinity-only model are highly correlated with the observed host and parasite degrees ($r = 0.95$; Figure SM 4a), but again the correlation is lower in the combined model ($r = 0.82$; Figure SM 4b), and further reduced with the phylogeny-only model ($r = 0.49$; Figure SM 4c). Therefore, predictions from the phylogeny model (and to a lesser extent, the combined model) are largely independent from study effort.

To contrast spatial predictions on the geographic distribution of missing links between the affinity and phylogeny models, we generated hotspot maps of undocumented host–parasite interactions. Maps were generated by summing the probabilities of undocumented interactions per host species, summing across host species per cell (0.5° resolution) using IUCN distributions terrestrial mammals (IUCN, 2019). To adjust for the uneven species richness of hosts, predictions per cell were standardized to a scale of 0–1. These hotspot maps of undocumented host–parasite interactions (Figure 4)

illustrate large differences in the spatial distribution of missing links. Closely mirroring the map of observed interactions (Figure SM 10), predictions from the affinity model (Figure 4a) highlight Europe as a hotspot of missing links, while those from the phylogeny model reveal highest density of missing links predicted in tropical and central America, followed by tropical Africa and Asia (Figure 4b). The hotspot map generated by the combined model was closer to that produced by the affinity-only model, but with higher relative risk in sub-Saharan Africa, south America and parts of south-east Asia (Figure 4c).

The top-ranked links from the affinity and combined models were largely dominated by humans and domesticated animal hosts, while the phylogeny models more often predicted links among wildlife (Figure 3) including endangered and relatively poorly studied species, some of which are critically endangered (IUCN, 2019). Parasites infecting large numbers and phylogenetic ranges of hosts were most often included in the top undocumented links in all models (e.g. *Rabies lyssavirus*, *Sarcoptes scabiei*, *Toxoplasma gondii* and *Trypanosoma cruzi*). These parasites are commonly cited as capable of causing disease in a large number of (and sometimes all) mammals, though the majority of which have not been directly investigated (Arlan & Morgan, 2017; Innes, 2010; Jansen et al., 2018; Rupprecht et al., 2008). The combined model included a larger diversity of parasite species among the top predicted links (Table S3).

By conducting targeted literature searches of the top predicted missing links for each model by data subset, we found multiple links

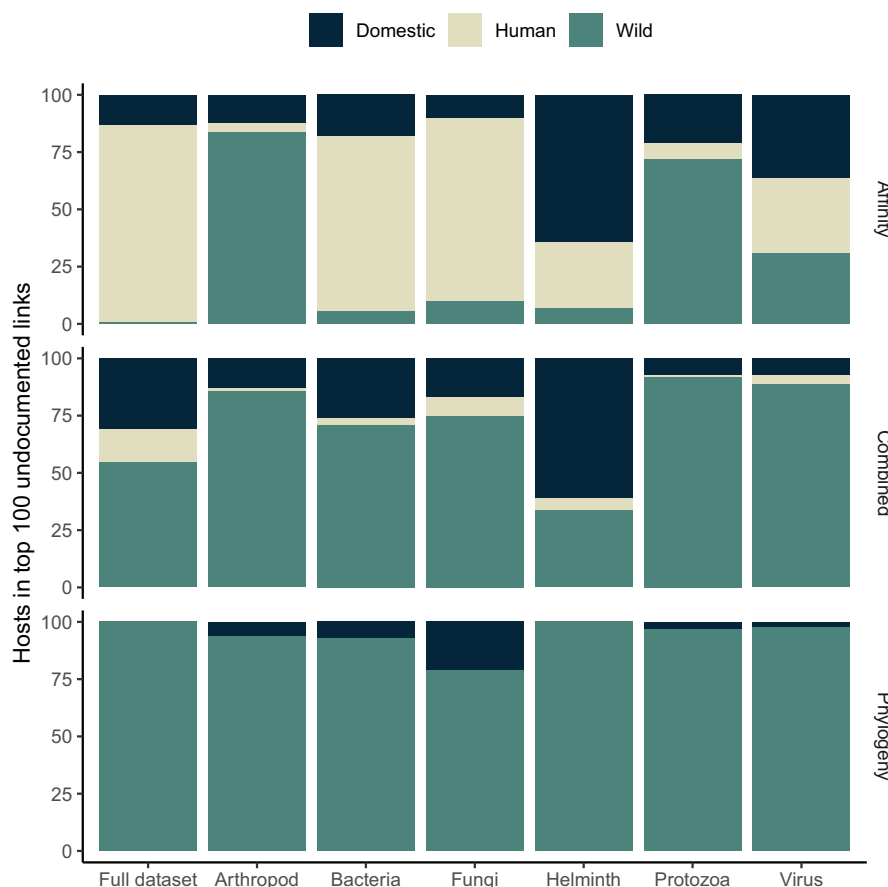
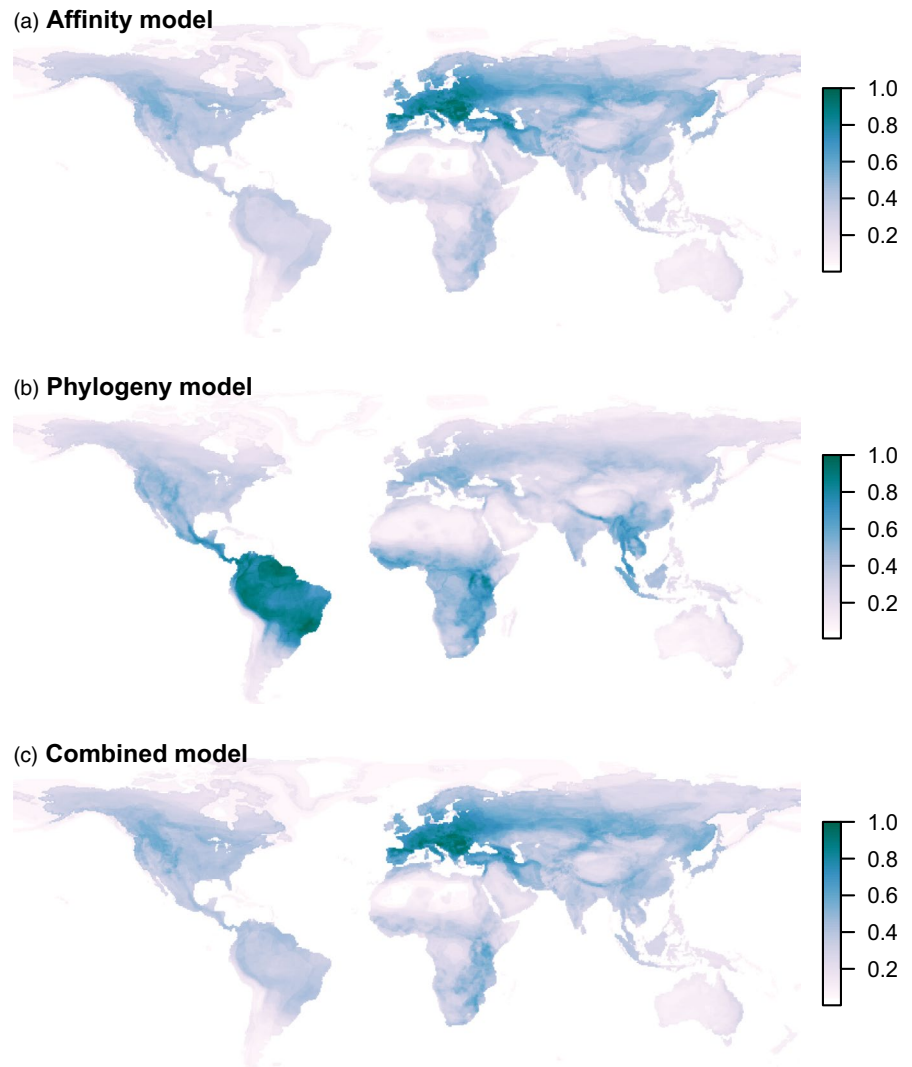


FIGURE 3 The frequencies of host types (human, domestic and wildlife) included in the top 100 predicted links per model and dataset combination, which were not documented in the original database. Plots are grouped by data subset as columns (the full dataset, and subsets by parasite type) and model as rows (affinity, combined and phylogeny)

FIGURE 4 Global hotspot maps showing the relative density of predicted but undocumented host–parasite interactions. Maps were generated by summing the probabilities of undocumented host–parasite interactions per host species, summing across host species per cell (0.5° resolution) with range maps for terrestrial mammals from the IUCN, then standardizing to a scale of 0–1. (a) Represents the relative undocumented link density based on predictions from the affinity model, while (b) represents the phylogeny model and (c) represents the combined model. Because these are sparse matrices, using average probabilities per cell or per host species would result in a map with all values close to zero. By taking the sum, we display information about relative risk and identify locations where additional sampling is more or less likely to reveal any previously undocumented interaction



had published support, but were not included in the original source databases (See SM for lists of top links and detailed results of literature searches). The combined model identified a greater number of links with literature support but which were not in the original database (46/90) compared to the phylogeny (39/90) and affinity (29/90) models (Table S3).

4 | DISCUSSION

Using a phylogeny-informed bipartite network model, we were able to predict missing links in a very large global mammal–parasite network. That we are able to make robust predictions, even with extremely sparse input data, indicates that this modelling approach may be useful in other large, data-poor ecological networks. We compared the performance of our joint model with an affinity-only (‘rich-get-richer’) model and a phylogeny-only model. While predictive performance in cross-validation was regularly higher for the affinity model (as measured by AUC), our literature searches showed the phylogeny model to be less prone to predicting ecologically

unlikely links, indicating that it may better capture the underlying ecological and evolutionary processes structuring host–parasite interactions. The layering of the affinity and phylogeny models allows the combined model to exploit the scale-free nature of many real-world networks while further correcting for ecologically unlikely links. The influence of this correction factor is evident in the lower predictive error of the combined model (Figure 2e,f; Table S1), its ability to segregate observed and unobserved interactions among its top predictions (Figure SM 3) and can be seen visually by contrasting the posterior interaction matrices (Figure SM 2).

Predictions from the affinity model tended to reflect existing degree distributions (SM 2.2), indicated by the dominance of humans and domesticated animals among the top predictions (Figure 3), and highlighting of Europe as a hotspot of undocumented links (Figure 4), likely reflecting the volume of research on well-studied taxa in this region. We might expect that predicted links among well-studied taxa would already be well-documented if common. However, affinity-based predictions may be important for large-scale public health initiatives if they implicate widespread or abundant species as reservoirs of emerging infections. The affinity model is also more

likely to identify parasite sharing among distantly related hosts, which has the potential to result in high mortality following host shifts (Farrell & Davies, 2019). While the affinity model predicted some links supported by our literature search, the top predictions included multiple links that are unlikely due to a mismatch in host ecology. For example, domestic cattle *Bos taurus* are predicted to be susceptible to infection by *Anisakis simplex*. *A. simplex* is a trophically transmitted nematode that uses aquatic mammals as final hosts, with marine invertebrates and fish as intermediate hosts (Buchmann & Mehrdana, 2016), implying that cattle may only be exposed to the parasite if fed a marine-based diet.

The phylogeny model uses only the evolutionary relationships among hosts to predict missing links, and was found to strongly mitigate the propagation of potential research biases compared to the affinity model (SM 2.2). The hotspots of predicted links from the phylogeny model show a spatial distribution in stark contrast to the affinity model, with highest density in tropical and central America, followed by tropical Africa and Asia (Figure 4). It is unsurprising that these understudied regions, with high host and parasite diversity, are identified as centres of undocumented host–parasite associations. The top undocumented links predicted by the phylogeny model also included fewer ecologically unlikely interactions, with only one unlikely interaction among those we investigated via literature review. *Echinococcus granulosus* is typically maintained by a domestic cycle of dogs eating raw livestock offal (Otero-Abad & Torgerson, 2013), and while wild canids such as Pampas fox *Lycalopex gymnocercus* are known hosts (Lucherini & Luengos Vidal, 2008), the phylogeny model predicts Hoary fox *Lycalopex vetulus* as a potential host. However, this interaction is unlikely as *L. vetulus* has a largely insectivorous diet (Dalponte, 2009), unlike the other members of its genus.

4.1 | Missing links

We identified a number of parasites which may impact the health of humans and domesticated animals. These include parasites currently considered a risk for zoonotic transmission such as *Alaria alata*, an intestinal parasite of wild canids—a concern as other *Alaria* species have been reported to cause fatal illness in humans (Murphy et al., 2012), and *Bovine viral diarrhoea virus 1*, which is not currently considered to be a human pathogen, but is highly mutable, has the ability to replicate in human cell lines, and has been isolated from humans on rare occasions (Walz et al., 2010). However, there is a large amount of effort that goes into studying infectious diseases of humans and domestic species, and it is likely that most contemporary associations among humans and described parasites have been recorded, even if not included in the aggregated databases because they occur rarely or are difficult to detect. For example, we predicted that humans could be infected by *Bartonella grahamii*, and found that the first recorded case was in an immunocompromised patient in 2013 (Oksi et al., 2013). Similarly, humans are predicted to be susceptible to *Mycoplasma haemofelis*, which was again reported in someone who was immunocompromised (dos Santos et al., 2008),

indicating that while these infections may pose little risk for a large portion of the human population, they are a serious concern for the health of immunocompromised individuals. These examples demonstrate our framework has the capacity to predict known human diseases, highlight parasites that are recognized zoonotic risks and identify a number of parasites that are currently unrecognized as zoonotic risks.

Applying link prediction methods to wildlife global host–parasite networks can additionally highlight both historical and contemporary disease threats to biodiversity (Farrell et al., 2021), and identify parasites with the potential to drive endangered species towards extinction. For example, the phylogeny models identified links reported only in literature from over 30 years ago, such as *T. cruzi* in the critically endangered cotton-top tamarin *Saguinus oedipus* (Marinkelle, 1982) and the vulnerable, black-crowned Central American squirrel monkey *Saimiri oerstedii* (Sousa, 1972). Our guided literature search also found evidence of severe infections in several endangered species such as rabies and sarcoptic mange *Sarcoptes scabiei* in Dhole *Cuon alpinus* (Durbin et al., 2005) and *Toxoplasma gondii* in critically endangered African wild dogs *Lycalopex pictus* which caused a fatal infection in a pup (Van Heerden et al., 1995). Our model also predicts that rabies and sarcoptic mange are likely to infect the endangered Darwin's fox *Lycalopex fulvipes*. Disease spread via contact with domestic dogs (notably *Canine distemper virus*) is currently one of the main threats to this species (Silva-Rodríguez et al., 2016). Considering that both rabies and sarcoptic mange from domestic dogs are implicated in the declines of other wild canids, they may pose a serious risk for the conservation of Darwin's foxes.

Because we did not include geographic constraints in our model, we may predict interactions among potentially compatible hosts and parasites that may be unrealized in nature due to lack of geographic overlap. These included *Trypanosoma cruzi*, which is currently restricted to the Americas (Browne et al., 2017), infecting endangered African species such as black rhinoceros *Diceros bicornis*, lowland gorilla *Gorilla gorilla* and chimpanzee *Pan troglodytes*. Although contemporary natural infections of chimpanzees by this parasite are unlikely due to geography, we found a report of a fatal infection of a captive individual in Texas (Bommineni et al., 2009). In addition to this example, our models identified multiple infections documented only in captive animals. While we found no published cases of natural infections, these demonstrate that the model is able to identify biologically plausible infection risks that are relevant for captive populations, and may present future risks in the face of host or parasite translocation or range shifts (Carlson, Albery, et al., 2021; Morales-Castilla et al., 2021).

4.2 | Iterative link prediction and parasite surveillance

Link prediction in host–parasite networks is a critical step in an iterative process of prediction and verification, whereby likely links are identified, queried and new links are added, allowing

predictions to be updated. For example, we identified a number of interactions that were first documented in the literature only after the source databases were assembled (e.g. *Nematodirus spathiger* in *Gazella leptoceros* (Said et al., 2018), *Toxoplasma gondii* in *Papio anubis* (Kamau et al., 2016) and *Trypanosoma cruzi* in horses (Bryan et al., 2016)), suggesting that prediction-guided literature searches offer a cost-effective solution to addressing knowledge gaps, maximizing the value of published literature and identifying targets for future field-based sampling. We view our effort as working towards the larger goal of expanding and filling out the global mammal–parasite network, which also includes programs for pathogen discovery, and field and collection-based parasite sampling.

We demonstrate that missing links in global databases of host–parasite interactions can be identified using information on known associations and the evolutionary relationships among host species. Link prediction represents a cost-effective approach for augmenting global databases used in the study of disease ecology and evolution. As we move down the list of most probable links, we will uncover links with infectious organisms that are less well-studied, but which may emerge as public or wildlife health burdens in the future. Through targeting research and surveillance efforts towards likely undocumented interactions, we can more efficiently gather baseline knowledge of the diversity of host–parasite interactions, ultimately supporting the development of fundamental theory in disease ecology and evolution (Stephens et al., 2016) and strengthening our understanding of disease spread and persistence in multi-host systems (Viana et al., 2014). In addition to prediction of host species that are susceptible or exposed to known infectious diseases, we can guide the proactive surveillance of multi-host parasites underlying contemporary disease burdens, and those which may emerge in the future (Carlson, Farrell, et al., 2021; Morales-Castilla et al., 2021).

4.3 | Future directions

In our study, all three models predicted links that were supported through targeted literature searches, though the host and parasite taxa differed. We therefore suggest that information on affinities and phylogeny are complementary, and while we place greater emphasis on the combined model here, choice of one model over another should depend on the goals of subsequent analyses, and whether it is important to minimize false positives or false negatives among predicted links.

Currently the approach is limited by the use of binary associations provided by current databases. These may reflect varying types of interactions, some of which are more important for global health and conservation. If available, these models may be fit with weighted rather than binary associations, allowing for modelling links as a function of prevalence, intensity of infection or to explicitly incorporate the amount of evidence supporting each link. In this way, sampling intensity may be directly incorporated into link

predictions, and help identify weakly supported interactions or sampling artefacts that may benefit from additional investigation.

Further, global interaction data are known to have sampling biases that cannot be easily adjusted for in models that directly or indirectly make predictions using the number of documented interactions per species. Our results show that incorporating phylogeny helps alleviate some of this bias, but so far are limited to using host phylogenies as well-resolved parasite phylogenies are often unavailable. However, the flexibility of the method allows for any information to be included, provided it can be represented as a distance matrix (Elmasri et al., 2020). Future models may be expanded to use trait or geographic distances among species to exclude ecological mismatches not currently captured by phylogeny, or re-formulated to add information on parasite phylogeny, if available. However, we caution that models with scaled distance matrices for both hosts and parasites may be challenging to fit.

We present our approach and predictions of top missing links here as a resource for future work on the macroecology of multi-host–multi-parasite dynamics. However, an important next step for data amalgamation and risk prediction is to move beyond binary associations, and quantify the nature of the association between host and parasite (Becker et al., 2020). In this way we may be able to predict not only the presence or absence of a particular host–parasite interaction, but the epidemiological role each host plays in parasite transmission (primary, intermediate or incidental host), the impact of infection on host fitness (Farrell & Davies, 2019) and better understand the ecologies of reservoir versus spillover hosts.

ACKNOWLEDGEMENTS

This work would not exist without the McGill Statistics-Biology Exchange (S-BEX) organized by Zofia Taranu, Amanda Winegardner and Russel Steele. The authors thank Will Pearse, Ignacio Morales-Castilla and Klaus Schliep for support and advice, and Colin Carlson and Greg Albery for friendly review of the manuscript. M.J.F. was supported by a Vanier NSERC CGS, the CIHR Systems Biology Training Program, the Quebec Centre for Biodiversity Science and the McGill Biology Department. M.E. was supported by the McGill University Department of Mathematics and Statistics, FRQNT and an NSERC PDF.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

AUTHORS' CONTRIBUTIONS

M.J.F. and T.J.D. designed the study; M.E., M.J.F., T.J.D. and D.A.S. designed the methods; M.J.F. compiled the data; M.J.F. and M.E. conducted the analyses; M.J.F. wrote the manuscript with input from T.J.D. All authors gave final approval for publication.

DATA AVAILABILITY STATEMENT

The amalgamated host–parasite matrix, results of all models, top predictions from each model and R scripts to format predictions and reproduce the figures can be found at doi: 10.6084/

m9.figshare.8969882 (Farrell, 2021). The R package used to fit the model described in Elmasri et al. (2020) is available at github.com/melmasri/HP-prediction.

ORCID

Maxwell J. Farrell  <https://orcid.org/0000-0003-0452-6993>

David A. Stephens  <https://orcid.org/0000-0001-9811-7140>

T. Jonathan Davies  <https://orcid.org/0000-0003-3318-5948>

REFERENCES

- Albery, G. F., Becker, D. J., Brierley, L., Brook, C. E., Christofferson, R. C., Cohen, L. E., Dallas, T. A., Eskew, E. A., Fagre, A., Farrell, M. J., Glennon, E., Guth, S., Joseph, M. B., Mollentze, N., Neely, B. A., Poisot, T., Rasmussen, A. L., Ryan, S. J., Seifert, S., ... Carlson, C. J. (2021). The science of the host-virus network. *Nature Microbiology*, 6, 1483–1492.
- Albery, G. F., Eskew, E. A., Ross, N., & Olival, K. J. (2020). Predicting the global mammalian viral sharing network using phylogeography. *Nature Communications*, 11, 2260.
- Arlian, L. G., & Morgan, M. S. (2017). A review of *Sarcoptes scabiei*: Past, present and future. *Parasites & Vectors*, 10, 297.
- Bartomeus, I., Gravel, D., Tylisanakis, J. M., Aizen, M. A., Dickie, I. A., & Bernard-Verdier, M. (2016). A common framework for identifying linkage rules across different types of interactions. *Functional Ecology*, 30, 1894–1903.
- Becker, D. J., Albery, G. F., Sjodin, A. R., Poisot, T., Bergner, L. M., Chen, B., Cohen, L. E., Dallas, T. A., Eskew, E. A., Fagre, A. C., Farrell, M. J., Guth, S., Han, B. A., Simmons, N. B., Stock, M., Teeling, E. C., & Carlson, C. J. (2022). Optimising predictive models to prioritise viral discovery in zoonotic reservoirs. *The Lancet Microbe*. [https://doi.org/10.1016/s2666-5247\(21\)00245-7](https://doi.org/10.1016/s2666-5247(21)00245-7)
- Becker, D. J., Seifert, S. N., & Carlson, C. J. (2020). Beyond infection: Integrating competence into reservoir host prediction. *Trends in Ecology & Evolution*, 35, 1062–1065.
- Bommineni, Y. R., Dick, E. J., Estep, J. S., Van de Berg, J. L., & Hubbard, G. B. (2009). Fatal acute Chagas disease in a chimpanzee. *Journal of Medical Primatology*, 38, 247–251.
- Braga, M. P., Razzolini, E., & Boeger, W. A. (2015). Drivers of parasite sharing among neotropical freshwater fishes. *Journal of Animal Ecology*, 84, 487–497.
- Browne, A. J., Guerra, C. A., Alves, R. V., da Costa, V. M., Wilson, A. L., Pigott, D. M., Hay, S. I., Lindsay, S. W., Golding, N., & Moyes, C. L. (2017). The contemporary distribution of *Trypanosoma cruzi* infection in humans, alternative hosts and vectors. *Scientific Data*, 4(1). <https://doi.org/10.1038/sdata.2017.50>
- Bryan, L. K., Hamer, S. A., Shaw, S., Curtis-Robles, R., Auckland, L. D., Hodo, C. L., Chaffin, K., & Rech, R. R. (2016). Chagas disease in a Texan horse with neurologic deficits. *Veterinary Parasitology*, 216, 13–17.
- Buchmann, K., & Mehrdana, F. (2016). Effects of anisakid nematodes *Anisakis simplex* (s.l.), *Pseudoterranova decipiens* (s.l.) and *Contracaecum osculatatum* (s.l.) on fish and consumer health. *Food and Waterborne Parasitology*, 4, 13–22.
- Carlson, C. J., Albery, G. F., Merow, C., Trisos, C. H., Zipfel, C. M., Eskew, E. A., Olival, K. J., Ross, N., & Bansal, S. (2021). Climate change will drive novel cross-species viral transmission. *bioRxiv*. <https://doi.org/10.1101/2020.01.24.918755>
- Carlson, C. J., Farrell, M. J., Grange, Z., Han, B. A., Mollentze, N., Phelan, A. L., Rasmussen, A. L., Albery, G. F., Bett, B., Brett-Major, D. M., Cohen, L. E., Dallas, T., Eskew, E. A., Fagre, A. C., Forbes, K. M., Gibb, R., Halabi, S., Hammer, C. C., Katz, R., ... Webala, P. W. (2021). The future of zoonotic risk prediction. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 376, 20200358.
- Carlson, C. J., Zipfel, C. M., Garnier, R., & Bansal, S. (2019). Global estimates of mammalian viral diversity accounting for host sharing. *Nature Ecology & Evolution*, 3, 1070–1075.
- Cleaveland, S., Laurenson, M. K., & Taylor, L. H. (2001). Diseases of humans and their domestic mammals: Pathogen characteristics, host range and the risk of emergence. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 356, 991–999.
- Clutton-Brock, J. (1999). *A natural history of domesticated mammals* (2nd ed.). Cambridge University Press.
- Dallas, T. (2016). helminthR: An R interface to the London Natural History Museum's Host-Parasite Database. *Ecography*, 39(4), 391–393. <https://doi.org/10.1111/ecog.02131>
- Dallas, T., Huang, S., Nunn, C., Park, A. W., & Drake, J. M. (2017). Estimating parasite host range. *Proceedings of the Royal Society B: Biological Sciences*, 284, 20171250.
- Dalponete, J. C. (2009). *Lycalopex vetulus* (Carnivora: Canidae). *Mammalian Species*, 847, 1–7.
- Davies, T. J., & Pedersen, A. B. (2008). Phylogeny and geography predict pathogen community similarity in wild primates and humans. *Proceedings of the Royal Society B: Biological Sciences*, 275, 1695–1701.
- dos Santos, A. P., dos Santos, R. P., Biondo, A. W., Dora, J. M., Goldani, L. Z., de Oliveira, S. T., de Sá Guimarães, A. M., Timenetsky, J., de Moraes, H. A., González, F. H., & Messick, J. B. (2008). Hemoplasma infection in HIV-positive patient, Brazil. *Emerging Infectious Diseases*, 14, 1922–1924.
- Durbin, L., Venkataraman, A., Hedges, S., & Duckworth, W. (2005). South Asia–south of the Himalaya (oriental). In C. Sillero-Zubiri, M. Hoffmann & D. Macdonald (Eds.), *Canids: Foxes, wolves, jackals and dogs* (chap. 8, pp. 210–219). Island Press.
- Elmasri, M., Farrell, M. J., Davies, T. J., & Stephens, D. A. (2020). A hierarchical Bayesian model for predicting ecological interactions using scaled evolutionary relationships. *The Annals of Applied Statistics*, 14(1). <https://doi.org/10.1214/19-aos1296>
- Ezenwa, V. O., Price, S. A., Altizer, S., Vitone, N. D., & Cook, K. C. (2006). Host traits and parasite species richness in even and odd-toed hoofed mammals, Artiodactyla and Perissodactyla. *Oikos*, 115, 526–536.
- Farrell, M. (2021). Data code: 'Predicting missing links in global host-parasite networks'. Figshare, <https://doi.org/10.6084/m9.figshare.8969882>
- Farrell, M. J., & Davies, T. J. (2019). Disease mortality in domesticated animals is predicted by host evolutionary relationships. *Proceedings of the National Academy of Sciences of the United States of America*, 116, 7911–7915.
- Farrell, M. J., Park, A. W., Cressler, C. E., Dallas, T., Huang, S., Mideo, N., Morales-Castilla, I., Davies, T. J., & Stephens, P. (2021). The ghost of hosts past: Impacts of host extinction on parasite specificity. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 376, 20200351.
- Fritz, S. A., Bininda-Emonds, O. R. P., & Purvis, A. (2009). Geographical variation in predictors of mammalian extinction risk: Big is bad, but only in the tropics. *Ecology Letters*, 12, 538–549.
- Gibson, D. I., Bray, R. A., & Harris, E. A. C. (2005). *Host-parasite database of the Natural History Museum, London*.
- Gomez, J. M., Nunn, C. L., & Verdu, M. (2013). Centrality in primate-parasite networks reveals the potential for the transmission of emerging infectious diseases to humans. *Proceedings of the National Academy of Sciences*, 110(19), 7738–7741. <https://doi.org/10.1073/pnas.1220716110>
- Grace, D., Gilbert, J., Randolph, T., & Kang'ethe, E. (2012). The multiple burdens of zoonotic disease and an ecohealth approach to their assessment. *Tropical Animal Health and Production*, 44, 67–73.
- Gravel, D., Poisot, T., Albouy, C., Velez, L., & Mouillot, D. (2013). Inferring food web structure from predator-prey body size relationships. *Methods in Ecology and Evolution*, 4, 1083–1090.

- Han, B. A., Schmidt, J. P., Alexander, L. W., Bowden, S. E., Hayman, D. T. S., & Drake, J. M. (2016). Undiscovered bat hosts of filoviruses. *PLoS Neglected Tropical Diseases*, 10, 1–10.
- Han, B. A., Schmidt, J. P., Bowden, S. E., & Drake, J. M. (2015). Rodent reservoirs of future zoonotic diseases. *Proceedings of the National Academy of Sciences of the United States of America*, 112, 7039–7044.
- Harmon, L. J., Losos, J. B., Jonathan Davies, T., Gillespie, R. G., Gittleman, J. L., Bryan Jennings, W., Kozak, K. H., McPeck, M. A., Moreno-Roark, F., Near, T. J., Purvis, A., Ricklefs, R. E., Schluter, D., Schulte, J. A., Seehausen, O., Sidlauskas, B. L., Torres-Carvajal, O., Weir, J. T., & Mooers, A. T. (2010). Early bursts of body size and shape evolution are rare in comparative data. *Evolution*, 64, 2385–2396.
- Huang, S., Bininda-Emonds, O. R. P., Stephens, P. R., Gittleman, J. L., & Altizer, S. (2014). Phylogenetically related and ecologically similar carnivores harbour similar parasite assemblages. *Journal of Animal Ecology*, 83, 671–680.
- Huang, S., Drake, J. M., Gittleman, J. L., & Altizer, S. (2015). Parasite diversity declines with host evolutionary distinctiveness: A global analysis of carnivores. *Evolution*, 69, 621–630.
- Innes, E. A. (2010). A brief history and overview of *Toxoplasma gondii*. *Zoonoses and Public Health*, 57, 1–7. <https://doi.org/10.1111/j.1863-2378.2009.01276.x>
- IUCN. (2019). *The IUCN red list of threatened species*.
- Jansen, A. M., Xavier, S. C. D. C., & Roque, A. L. R. (2018). Trypanosoma cruzi transmission in the wild and its most important reservoir hosts in Brazil. *Parasites & Vectors*, 11, 502.
- Johnson, J. M., & Khoshgoftaar, T. M. (2019). Survey on deep learning with class imbalance. *Journal of Big Data*, 6(1), 1–19. <https://doi.org/10.1186/s40537-019-0192-5>
- Kamau, D., Kagira, J., Maina, N., Mutura, S., Mokua, J., & Karanja, S. (2016). Detection of natural *Toxoplasma gondii* infection in olive baboons (*Papio anubis*) in Kenya using nested PCR. The 11th Kkuat Scientific, Technological and Industrialization Conference and Exhibitions Conference Proceedings.
- Kamiya, T., O'Dwyer, K., Nakagawa, S., & Poulin, R. (2014). What determines species richness of parasitic organisms? A meta-analysis across animal, plant and fungal hosts. *Biological Reviews of the Cambridge Philosophical Society*, 89, 123–134.
- Lindfors, P., Nunn, C. L., Jones, K. E., Cunningham, A. A., Sechrest, W., & Gittleman, J. L. (2007). Parasite species richness in carnivores: Effects of host body mass, latitude, geographical range and population density. *Global Ecology and Biogeography*, 16, 496–509.
- Lucherini, M., & Luengos Vidal, E. M. (2008). *Lycalopex gymnocercus* (Carnivora: Canidae). *Mammalian Species*, 820, 1–9.
- Luis, A. D., O'Shea, T. J., Hayman, D. T. S., Wood, J. L. N., Cunningham, A. A., Gilbert, A. T., Mills, J. N., & Webb, C. T. (2015). Network analysis of host–virus communities in bats and rodents reveals determinants of cross-species transmission. *Ecology Letters*, 18(11), 1153–1162. <https://doi.org/10.1111/ele.12491>
- Marinkelle, C. J. (1982). The prevalence of *Trypanosoma (Schizotrypanum) cruzi* infection in Colombian monkeys and marmosets. *Annals of Tropical Medicine & Parasitology* 76, 121–124. PMID: 6807227.
- Morales-Castilla, I., Matias, M. G., Gravel, D., & Araújo, M. B. (2015). Inferring biotic interactions from proxies. *Trends in Ecology & Evolution*, 30, 347–356.
- Morales-Castilla, I., Pappalardo, P., Farrell, M. J., Aguirre, A. A., Huang, S., Gehman, A. L. M., Dallas, T., Gravel, D., & Davies, T. J. (2021). Forecasting parasite sharing under climate change. *Philosophical transactions of the Royal Society B: Biological Sciences*, 376, 20200360.
- Murphy, T. M., O'Connell, J., Berzano, M., Dold, C., Keegan, J. D., McCann, A., Murphy, D., & Holden, N. M. (2012). The prevalence and distribution of *Alaria alata*, a potential zoonotic parasite, in foxes in Ireland. *Parasitology Research*, 111, 283–290.
- Nunn, C. L., Altizer, S., Jones, K. E., & Sechrest, W. (2003). Comparative tests of parasite species richness in primates. *The American Naturalist*, 162, 597–614.
- Oksi, J., Rantala, S., Kilpinen, S., Silvennoinen, R., Vornanen, M., Veikkolainen, V., Eerola, E., & Pulliainen, A. T. (2013). Cat scratch disease caused by *Bartonella grahamii* in an immunocompromised patient. *Journal of Clinical Microbiology*, 51, 2781–2784.
- Olival, K. J., Hosseini, P. R., Zambrana-Torrel, C., Ross, N., Bogich, T. L., & Daszak, P. (2017). Host and viral traits predict zoonotic spillover from mammals. *Nature*, 546, 646–650.
- Otero-Abad, B., & Torgerson, P. R. (2013). A systematic review of the epidemiology of echinococcosis in domestic and wild animals. *PLoS Neglected Tropical Diseases*, 7, e2249.
- Page, R. D. M. (1993). Parasites, phylogeny and cospeciation. *International Journal for Parasitology*, 23(4), 499–506. [https://doi.org/10.1016/0020-7519\(93\)90039-2](https://doi.org/10.1016/0020-7519(93)90039-2)
- Pandit, P. S., Doyle, M. M., Smart, K. M., Young, C. C. W., Drape, G. W., & Johnson, C. K. (2018). Predicting wildlife reservoirs and global vulnerability to zoonotic flaviviruses. *Nature Communications*, 9, 5425.
- Park, A. W. (2019). Food web structure selects for parasite host range. *Proceedings of the Royal Society B: Biological Sciences*, 286, 20191277.
- Park, A. W., Farrell, M. J., Schmidt, J. P., Huang, S., Dallas, T. A., Pappalardo, P., Drake, J. M., Stephens, P. R., Poulin, R., Nunn, C. L., & Davies, T. J. (2018). Characterizing the phylogenetic specialism-generalism spectrum of mammal parasites. *Proceedings of the Royal Society B: Biological Sciences*, 285, 20172613.
- Pilosof, S., Morand, S., Krasnov, B. R., & Nunn, C. L. (2015). Potential parasite transmission in multi-host networks based on parasite sharing. *PLoS ONE*, 10, e0117909.
- Ricci, F., Rokach, L., & Shapira, B. (2011). *Introduction to recommender systems handbook*. Springer.
- Rupprecht, C. E., Barrett, J., Briggs, D., Cliquet, F., Fooks, A. R., Lumlertdacha, B., Meslin, F. X., Müller, T., Nel, L. H., Schneider, C., Tordo, N., & Wandeler, A. I. (2008). Can rabies be eradicated. *Developmental Biology*, 131, 95–121.
- Said, Y., Gharbi, M., Mhadhbi, M., Dhibi, M., & Lahmar, S. (2018). Molecular identification of parasitic nematodes (Nematoda: Strongylida) in feces of wild ruminants from Tunisia. *Parasitology*, 145, 901–911.
- Shwab, E. K., Saraf, P., Zhu, X. Q., Zhou, D. H., McFerrin, B. M., Ajzenberg, D., Schares, G., Hammond-Aryee, K., van Helden, P., Higgins, S. A., Gerhold, R. W., Rosenthal, B. M., Zhao, X., Dubey, J. P., & Su, C. (2018). Human impact on the diversity and virulence of the ubiquitous zoonotic parasite *Toxoplasma gondii*. *Proceedings of the National Academy of Sciences of the United States of America*, 115, 201722202.
- Silva-Rodríguez, E., Farias, A., Moreira-Arce, D., Cabello, J., Hidalgo-Hermoso, E., Lucherini, M., & Jiménez, J. (2016). *Lycalopex fulvipes*. The IUCN Red List of Threatened Species 2016.
- Smith, K. F., Acevedo-Whitehouse, K., & Pedersen, A. B. (2009). The role of infectious diseases in biological conservation. *Animal Conservation*, 12, 1–12.
- Sousa, O. E. (1972). Anotaciones sobre la enfermedad de Chagas en Panamá. Frecuencia y distribución de *Trypanosoma cruzi* y *Trypanosoma rangeli*. *Revista de Biología Tropical*, 20, 167–179.
- Stephens, P. R., Altizer, S., Smith, K. F., Alonso Aguirre, A., Brown, J. H., Budischak, S. A., Byers, J. E., Dallas, T. A., Jonathan Davies, T., Drake, J. M., Ezenwa, V. O., Farrell, M. J., Gittleman, J. L., Han, B. A., Huang, S., Hutchinson, R. A., Johnson, P., Nunn, C. L., Onstad, D., ... Poulin, R. (2016). The macroecology of infectious diseases: A new perspective on global-scale drivers of pathogen distributions and impacts. *Ecology Letters*, 19, 1159–1171.
- Stephens, P. R., Pappalardo, P., Huang, S., Byers, J. E., Critchlow, R., Farrell, M. J., Gehman, A., Ghai, R. R., Haas, S., Han, B. A., Park, A. W., Schmidt, J. P., Altizer, S., Ezenwa, V. O., & Nunn, C. L. (2017). Global mammal parasite database version 2.0. *Ecology*, 98, 42–49.

- Taylor, L. H., Latham, S. M., & Woolhouse, M. E. J. (2001). Risk factors for human disease emergence. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 356(1411), 983–989. <https://doi.org/10.1098/rstb.2001.0888>
- Van Heerden, J., Mills, M. G., Van Vuuren, M. J., Kelly, P. J., & Dreyer, M. J. (1995). An investigation into the health status and diseases of wild dogs (*Lycaon pictus*) in the Kruger National Park. *Journal of the South African Veterinary Association*, 66, 18–27.
- Viana, M., Mancy, R., Biek, R., Cleaveland, S., Cross, P. C., Lloyd-Smith, J. O., & Haydon, D. T. (2014). Assembling evidence for identifying reservoirs of infection. *Trends in Ecology & Evolution*, 29, 270–279.
- Walz, P. H., Grooms, D. L., Passler, T., Ridpath, J. F., Tremblay, R., Step, D. L., Callan, R. J., & Givens, M. D. (2010). Control of bovine viral diarrhoea virus in ruminants. *Journal of Veterinary Internal Medicine*, 24, 476–486.
- 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, 150049.
- Wardeh, M., Sharkey, K. J., & Baylis, M. (2020). Integration of shared-pathogen networks and machine learning reveals the key aspects of zoonoses and predicts mammalian reservoirs. *Proceedings of the Royal Society B: Biological Sciences*, 287, 20192882.
- Watts, D. J., & Strogatz, S. H. (1998). Collective dynamics of 'small-world' networks. *Nature*, 393(6684), 440–442. <https://doi.org/10.1038/30918>
- Wiens, J. J., Ackerly, D. D., Allen, A. P., Anacker, B. L., Buckley, L. B., Cornell, H. V., Damschen, E. I., Jonathan Davies, T., Grytnes, J. A., Harrison, S. P., Hawkins, B. A., Holt, R. D., McCain, C. M., & Stephens, P. R. (2010). Niche conservatism as an emerging principle in ecology and conservation biology. *Ecology Letters*, 13, 1310–1324.
- Wilson, D. E., & Reeder, D. M. (2005). *Mammal species of the world: A taxonomic and geographic reference* (3rd ed.). Johns Hopkins University Press.

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.

How to cite this article: Farrell, M. J., Elmasri, M., Stephens, D. A., & Davies, T. J. (2022). Predicting missing links in global host–parasite networks. *Journal of Animal Ecology*, 00, 1–12. <https://doi.org/10.1111/1365-2656.13666>