Predicting Missing Virus-Host Links for Avian Hosts using Graph Neural Networks

Akshath Shvetang Anna
aanna7@gatech.edu
Georgia Institute of Technology
Atlanta, Georgia, USA

Grace Driskill
gdriskill3@gatech.edu
Georgia Institute of Technology
Atlanta, Georgia, USA

Julia Qian
jqian310@gatech.edu
Georgia Institute of Technology
Atlanta, Georgia, USA

Abstract

A virus's host range is an important indicator of its zoonotic potential. Predicting the missing hosts of viruses is thus an important step in supporting zoonotic disease surveillance, outbreak prevention and outbreak response. Here, we survey current methods and approaches in predicting missing links between animal species and pathogens. We then create an aggregated dataset of avian-virus interaction features, including avian-virus interactions, bird features, and virus features. We apply a graph neural network model to predict associations between bird species and viruses, using an auto-encoded embedding of the novel avian-virus dataset. The same technique was used on an existing mammalian-virus dataset. For each virus-bird pair, the average distance of the host to the known hosts of the virus and the average distance of the host to the rest of the avian species are calculated. The predictions of missing avian virus links were then cross-checked with existing academic papers and news articles for validity.

ACM Reference Format:
Akshath Shvetang Anna, Grace Driskill, and Julia Qian. 2024. Predicting Missing Virus-Host Links for Avian Hosts using Graph Neural Networks. In *CSE 8803 EPI, Fall, 2024, Atlanta, GA.* ACM, New York, NY, USA, 7 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

A zoonotic disease is an infectious illness transmitted from animals to humans. Approximately 60% of all human illnesses originate from animals [16], including high-profile diseases such as Ebola and COVID-19 (both believed to originate from bats). The WHO estimates that approximately 1 billion cases of zoonoses occur annually, with millions of deaths [27]. Avian zoonoses, originating from birds, present a unique challenge regarding surveillance and predictions. Birds, like mammals, can act as hosts to viruses. Alternatively, they can spread the disease to vectors which then pass the infection to humans (such as the bird-mosquito-human pathway for West Nile virus). Moreover, birds can greatly increase the geographic range of a disease by spreading it anywhere along their migration routes [29], infecting both humans and domesticated animals.

CSE 8803 EPI, November 2024, Atlanta, GA, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM
https://doi.org/XXXXXXX.XXXXXXX

Spillover is the event in which a disease isolated to one species begins infecting another. The range of species a virus can infect (host plasticity) is directly related to its potential to spillover to humans and its geographic reach [21]. Predicting a virus's host range can grant insight into which disease may become zoonotic, helping forecast potential outbreaks and supporting surveillance efforts.

Research efforts in this area by Wardeh et al. [35] have leveraged machine learning to predict previously unknown connections between mammal species and known viruses. Their results reveal over 20,000 potential unknown associations between viruses and mammals, pointing to how small a fraction of virus-mammal associations are known to researchers. Although their model provides extensive predictions for mammalian hosts, with 1.67 million undiscovered viruses that exist in mammals and bird hosts [6], this remains an incomplete picture.

Host-virus interactions, as well as ecological interactions as a whole, can be represented through biological networks [20]. Increasingly, machine learning techniques, in particular neural networks, have been applied to such networks as a means to analyze and identify patterns within data. Common tasks include node or edge classification, node clustering, link prediction, and recommendation [36]. Within the area of zoonotic diseases, graph neural networks (GNNs) have been leveraged to predict viral hosts, typically for prokaryotic viruses [31]. Virus-host predictions can be interpreted as a link prediction problem, where predictions are made about the probability of an edge forming between a host and a virus node.

This project aims to apply GNNs to analyze host-pathogen and interspecies avian datasets and predict unknown associations between birds and viruses. This will grant better insight into a virus's range of hosts and by extension their potential to spill over to humans.

2 Response to Milestone Comments

We have made additional progress for the final report to complete our stated goals.

A future extension for the project is the application of transfer learning between the mammalian dataset and the avian dataset. The pre-existing dataset is much more robust than our constructed avian dataset. It contains more features such as details about each mammals' habitat as well as specifics on different viruses. Training a model on either the mammalian dataset or the mammalian viruses could provide more robust insight into how similar viruses could interact with birds. For now we chose to focus on directly applying the GNN training process to the avian data since we were primarily interested in exploring whether a GNN model can accomplish this task. Additionally, the model is not resource-intensive to train, so

it is straightforward to use it directly. However training it well on the mammalian dataset first and then applying that to the avian dataset could be an interesting future step to see if it improves the quality of the predictions for the avian data.

3 Survey of Related Work

Given the global health burden presented by zoonotic disease and the relationship between a virus's host range and its zoonotic potential [21], there has been an increasing interest in understanding the full scope of a virus's existing and potential hosts.

Wardeh et al. [35] aimed to predict unknown associations between mammals and viruses through ensemble learning. They created three machine learning models to predict virus-mammal associations, mammal-virus associations, and missing edges in a global bipartite network, before applying ensemble voting. Although their paper is focused on mammal-virus associations, their goal to predict unknown animal-virus associations provides insight into how an assessment of avian-virus associations can be approached. One limitation of their approach is that, by using interaction data, their results may be biased due to targeted research. Associations may be over or under-reported due to factors like geographical accessibility (i.e. it may be harder to reach some populations of animals for testing), or an increased interest in certain mammals due to current outbreaks. Their global network view provides an alternative method to deal with these research and sampling biases.

Olival et al. [26] investigate factors that impact a virus's host range and transmissibility to other species. They discover that viral richness, the number of different viruses found in a host species, is impacted by research effort, mammal sympatry (when two species live in the same area) and host geographic range. In turn, factors that impact a mammal species's zoonotic viral richness include overall viral richness, phylogenetic distance from humans, host order and research effort. Although their research targets human-virus associations, their research highlights important geographical and phylogenetic factors that impact a virus's transmission from one species to another. We can apply the features they found that impact viral richness to our model to better estimate how many viruses different avian-species host. Similar to the previous paper, their work is also potentially biased by the attention each virus receives in the research space and cannot account for associations that currently remain undiscovered.

Poisot et al. [28] address the challenge of predicting missing host-virus interactions as a network problem. They represent the network as a bipartite structure, where virus nodes connect to host nodes where an association is known. They used a combined linear filtering and singular value decomposition model with hyper-parameters for determining the importance of different network structure characteristics. By choosing hyper-parameters that emphasize the network's connectance instead of in-degree or out-degree, they helped to reduce the bias caused by uneven research efforts, where certain parts of the network have more data due to concentrated study. Their techniques lead to predictions for possible new host-virus interactions that are robust to these data biases and suggest that future efforts for virus discovery be focused on the Amazon Basin and sub-Saharan Africa. While their method effectively handled the issue of incomplete and unbalanced data, a limitation was the exclusion of host and viral traits as predictors, which could enhance the model's performance.

Altizer et al. [1] investigates the relationship between animal migration and zoonotic diseases, suggesting that migration can both facilitate and mitigate disease spreading. A common concern is that migratory species may move diseases long distances geographically, and although this is a possibility, there are few documented examples of this phenomenon. More likely concerns are that the migratory species themselves are exposed to a wider range of diseases throughout their migration and that hotspots for diseases are created in places where multiple species of migratory animals stop to rest. An example of this is Delaware Bay, where stopping shorebirds create a hotspot of avian influenza. The stress of migration may lower the animals' immune system making them more susceptible to disease. However, Altizer et al. [1] also suggest that migration can actually decrease the spread of disease because staying in one habitat for a long period of time allows parasites to accumulate. Additionally, migration removes diseased members of a species from a community, and then decreases the chances of them spreading the illness, because they are not able to complete the migration.

This information is important when considering what factors to take into account when trying to predict potential zoonotic diseases. Due to the complicated role that migration has in zoonotic disease transmission, including migratory animals and information on their migrations could impact the model.

Shang and Sun [32] studied a similar problem of predicting hosts of prokaryotic viruses, which is relevant in the microbiology field. They approached host prediction as a link prediction task in a knowledge graph created from protein and DNA sequence features of the involved viruses and hosts, such as protein organization and sequence similarity. Their model has an encoder-decoder structure with a graph convolution network (GCN) encoder to utilize the topological structure and embedded node features from the graph. The decoder estimates the probability of a virus-prokaryote pair having a link. This study reports a sizable improvement in host prediction accuracy by using this model, from 43% to 80%. This suggests that GNNs and similar graph-based architecture could be promising for predicting virus-host interactions for zoonotic disease too. Additionally, this study's model only uses sequence data without integrating other traits of the hosts or viruses, which is a potential place for improvement.

Similarly, Du et al. [9] applied GNNs to predict phage-host interactions, with the intent to overcome the constraints that exist on current prediction models caused by the number of potential unknown interactions in comparison to the datasets available. They construct a heterogeneous phage-prokaryote graph using virus-virus protein and virus-host DNA similarities. They then leveraged a light graph convolutional network (LGCN) to recommend the host with the highest probability for each virus and achieved results competitive with existing virus-host interaction prediction methods. However, their constructed network did not include host-host interactions, which could improve the quality and accuracy of their predictions.

4 Problem Definition

The problem we aim to address is that, while some associations between hosts and viruses are known, many associations between hosts and known viruses remain unknown. Using a dataset of bird traits, a dataset of virus traits and a dataset of bird-viruses associations, our goal is to extract key features from each to predict these missing associations between birds and viruses.

We will treat this as a link-prediction problem and apply GNNs to predict links between viral and avian nodes. We will first apply our GNN model to the Divide-and-conquer dataset of mammal-virus interactions [35]. This will allow us to evaluate the performance of our approach against the predictions made by Wardeh et al. [35]'s three-perspective approach using a robust dataset containing phylogenetic, ecological, and genetic traits. Then, we will apply our GNN model to our own constructed dataset of avian-virus interactions containing a smaller feature set. In other words, we will predict which birds are potential hosts to known viruses, with an emphasis on predicting previously unreported relationships.

5 Method

5.1 Data

For this project, we plan to use the datasets used in Wardeh et al. [35] for information on mammal-virus interactions, mammal traits and mammalian viruses. Additionally, we are creating additional datasets on virus-avian interactions, avian traits and avian viruses.

5.1.1 Virus-Host Interactions. Virus-bird interactions were downloaded from the NCBI Virus Database [4]. 4743 unique virus-bird interactions were found. This included 1184 unique avian hosts and 1929 unique viruses.

5.1.2 Avian Traits. A dataset of avian hosts identified in virus-host interactions, along with their traits, was constructed. The host Taxon ID and species lineage (order, family, and genus) were obtained from the NCBI Taxonomy Database using NCBI's web API, Entrez Programming Utilities [3]. The traits for each bird include phylogenetic, ecological, and geospatial information, mirroring the mammal traits used in Wardeh et al. [35]. The following sources are used for this: AVONET [33] for information on birds' body mass, habitat, range, migration, trophic level and niche, and primary lifestyle, and BirdTree [15] for birds' phylogeny and taxonomy.

1,184 unique avian hosts were identified in the virus-host interactions. However, 357 bird scientific names were not found in AVONET. Some were due to birds' scientific name being at the subspecies level, for which traits at the species level were substituted. Some were due to former species names in either AVONET or NCBI, for which traits for the synonymous scientific name were used. Species names were cross-checked on Avibase [24] to ensure accuracy. These efforts allowed 118 more species to be found in AVONET. Therefore, 945 unique avian hosts from the virus-host interactions dataset had sufficient data in AVONET.

A phylogenetic tree for these species was downloaded from BirdTree. This tree was used to calculate a measure of evolutionary distinctiveness of the hosts by summing the lengths of the branches leading to each species.

Feature	Description
Scientific Name	Host's scientific name
NCBI Taxon ID	Taxid as given by NCBI Taxonomy, used as ID for host
Family	Host family
Genus	Host Genus
Mass	Species average body mass in grams
Trophic Level	The position the species occupies in a food web. Herbivore, Carnivore, Scavenger or Omnivore
Trophic Niche	A more specific categorization of the species' diet. Frugivore, Granivore, Vertivore, Aquatic Predator, etc.
Primary Lifestyle	Where the species spends the majority of its time. Aerial, Terrestrial, Insessorial, Aquatic, or Generalist
Range Size	The total area of the mapped range of the species in square kilometers
Habitat	The environment of the species. Desert, Grassland, Forest, Human modified, Marine, etc.
Habitat Density	The species density of this species' habitat. Discrete measure 1 to 3.
Migration	Categorizes this species' migration habits as 1 = Sedentary, 2= Partially migratory or 3 = Migratory
Evolutionary Distinctiveness	Measures how isolated a species is on a phylogenetic tree

Table 1: Avian traits.

5.1.3 Viral Traits. We also built a dataset for avian viruses and their traits. For each virus species in the virus-avian interaction dataset, the official virus species name is taken from the NCBI Virus Database [4] and ICTV [23]. The virus Taxon ID and species lineage (order, family, and genus) are obtained from the NCBI Taxonomy Database using NCBI's web API Entrez Programming Utilities [3].

The viral genome traits are obtained from ViralZone [13] and ICTV. These include the viral genome composition (1=RNA, 0=DNA), if the virus has double-stranded genetic material (1=double stranded, 0=single stranded), cytoplasmic replication (1=yes, 0=no), if the virus is enveloped (1=yes, 0=no), if the genome is circular (1=circular, 0=linear), if the virus is segmented or monopartite (1=segmented, 0=monopartite), and if the genetic material is positive or negative sense.

Genome-specific information is also included. Genomic size (in KB), GC content, and number of coding segments are taken from the NCBI Genome Assembly database [8].

Both ViralZone and the ICTV database are used to get information on replication sites, viral release mechanisms, and methods of cell entry.

Traits are assigned for each virus species. Species without data on certain traits are assigned traits one level up their lineage (i.e. genus, then order, then family).

Overall, there were 1795 different avian viruses at a species level that we were able to find at least one trait.

Feature	Description
Species	Species name
NCBI Taxon ID	Taxid as given by NCBI Taxonomy, used as ID for virus
Family	Virus family
Genus	Virus Genus
Envelope	1 if virus has an envelope, 0 if not
Circular	1 if the viral genome is circular, 0 if it is linear
Double Stranded	1 if the genome is double-stranded, 0 if not
RNA	1 if the viral genome is RNA, 0 if it is DNA
Segmented	1 if the genome is segmented, 0 if it is monopartite
Positive Sense	1 if the RNA is positive-sense, 0 otherwise or if genome is DNA
Negative Sense	1 if the RNA is negative-sense, 0 otherwise or if genome is DNA
Size	Size of genome in KB
GC	Guanine-Cytosine content of genome
Genes	Number of genes
Budding	Virus is released through budding
Lysis	Virus is released through lysis
Release Other	Virus is released through other means or is unspecified
Cytoplasm	1 if gene transcription occurs in the cytoplasm, 0 if it occurs in the nucleus
CE Clathrin	Viral entry through clathrin-mediated endocytosis
CE Receptor	Viral entry using receptors
CE Glycoprotein	Viral entry using glycoproteins
CE Other	Viral entry though other or unspecified means

Table 2: Virus traits.

5.1.4 Encoding Categorical Features. Both the avian and virus traits dataset contains mainly categorical data, so the choice of how to encode these features could have a significant impact on the results. One hot encoding creates a new binary feature for each category of the feature. This significantly increases the dimensionality of the dataset, which can lead to overfitting and increased computational complexity. Ordinal encoding, assigning an integer value for each category, preserves the number of dimensions, but assigns an order on the categories although they may not have a meaningful order. To overcome these issues, Dfencoder [18] was used to encode the host and virus traits. Dfencoder normalizes continuous variables and creates a dense embedding of categorical variables using an autoencoder.

5.2 Model

A bipartite graph was constructed using the virus-host interaction dataset, where the nodes represent either virus species or host species, and an edge represents a documented interaction between a virus and a host (i.e., the host has been infected by the virus). The host and bird traits were encoded using dfencoder; the encoded traits then served as the node features.

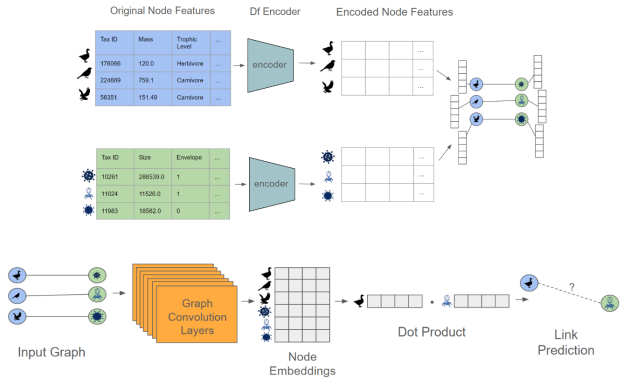

Figure 1: A Dfencoder was used to create encoded node features from the original viral/host traits. These were then passed to a GNN model to evaluate the likelihood of links.

This bipartite graph was then input into a Graph Neural Network (GNN) model. The PyTorch Geometric library [11] and its resources were used for implementing and training the model. The GNN model used is comprised of multiple SageConv convolution layers (GraphSAGE) [12]. Each SageConv layer aggregates information from a node's neighbors, iteratively refining node representations by combining local node features with the features of neighboring nodes, thus capturing higher-order relational patterns and graph structure. The use of multiple SageConv layers allows for information to propagate further through the graph as it spread neighbor-to-neighbor. The final output of the model is learned node embeddings for each host and virus node.

To predict potential interactions between a host and a virus, the dot product of the embeddings of a host and a virus node is computed. This dot product provides an edge score that represents the likelihood of an interaction between a given host and virus, allowing for suggestions of previously unobserved interactions in the bipartite graph.

6 Experiments/Results

Our experiments were designed to answer the following questions:

- Can this model predict known interactions for both the mammalian and avian datasets when known interactions are withheld from training?
- Do the previously unobserved interactions that this model predicts for the avian dataset make sense after doing a manual analysis of them?

6.0.1 Training Procedure. The dataset was split into train, validation, and test sets by randomly splitting the edges of the model with a negative sampling ratio of 2. All nodes were present across the three sets, with only edges being restricted. Thus, this was a transductive, as opposed to an inductive, learning setup, which can have better testing accuracy, at the expense of generalizability Lachaud et al. [22]. The model was trained through mini-batch gradient descent with the Adam optimizer [17] and a binary cross entropy loss. Negative sampling is included to discourage the model from

overly predicting new links, so that the ones that are still newly predicted despite that are ones of greater interest.

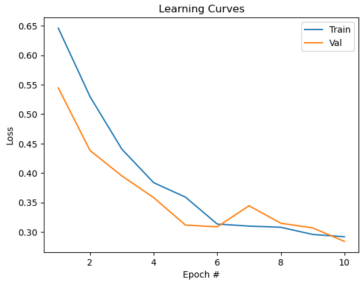

Figure 2: Loss Curves of Model Training on Avian Dataset

6.0.2 Testing against Held-out Test Set. Inherently, there is a class-imbalance between positive instances (edges) and negative instances (no edge). While Area Under the Receiver Operating Characteristic Curve (AUROC) is often chosen for its robustness to class imbalance as am evaluation metric, the Area Under the Precision-Recall Curve (AUPR) is better suited for link prediction tasks and better conveys the difficulty of the classification problem [25]. As seen in Figure 3, the model scored a relatively high average precision, which is an approximation for the area under the precision recall curve [2].

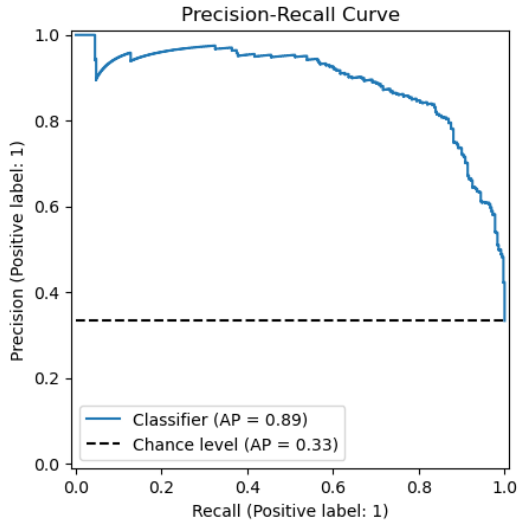

Figure 3: The Precision-Recall Curve on the Avian Validation Set

6.0.3 Use with Mammalian Dataset from Wardeh et al. [35]. As a plausibility check, this exact pipeline was applied to a further cleaned version of the mammalian-virus data from Wardeh et al. [35]. The model showed a similarly high AP on this dataset as well as can be seen in Figure 4. Further analysis can be done to compare the new links that are predicted from the GNN model to those

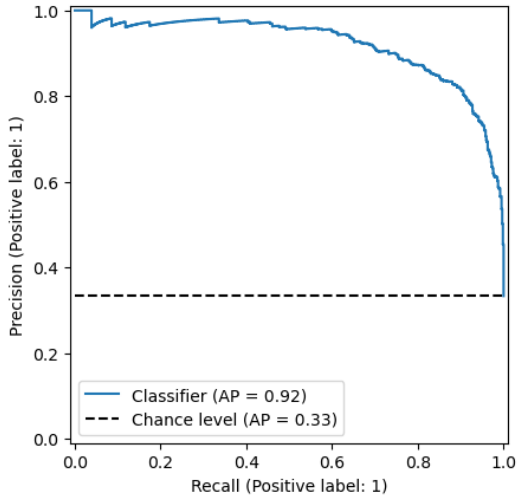

Figure 4: The Precision-Recall Curve on the Mammalian Validation Set

predicted by the different three-perspective based approach used by Wardeh et al. [35]. Fine tuning the GNN model to more closely resemble the additional predictions from their already-studied approach can better ensure that the predictions from the model are reasonable and relevant. If the GNN model is able to achieve similar predictions, it would provide a more streamlined and simpler alternative, compared to the training of an ensemble of localized models for each and every individual virus and host in the dataset as Wardeh et al. [35] did.

6.0.4 Manual Analysis of Newly Predicted Interactions. A manual analysis of the top five most likely predicted interactions previously unobserved was performed to investigate their viability. First, the phylogenetic distances of the avian hosts was considered as viruses often infect closely related hosts [14]. Using the same phylogenetic tree used to calculate the evolutionary distinctiveness of the avian species, the pairwise phylogenetic distances was calculated for each pair of species. Then, for the novel predicted interactions, the average distance of the host to the known hosts of the virus and the average distance of the host to the rest of the avian species are calculated.

For three of the top five predicted interactions, the host had a shorter average phylogenetic distance to known hosts of the virus than distance to other species in the dataset. This indicates that these hosts are evolutionarily similar, supporting that these are reasonable new virus-host interactions.

The top predicted interaction is between Myiopsitta monachus (monk parakeet) and Avipoxvirus. Myiopsitta monachus belongs to the family Psittacidae (parrots). Avipoxvirus is a highly prevalent disease among birds and has been recorded to have its own strain among Psittacidae, known as Psittacinepox virus [5]. A 2015 study found that avipoxvirus has been recorded in other parakeet species, in an outbreak at a conservation facility that affected parrots, parakeets, and macaws [10]. Given that Myiopsitta monachus are considered highly invasive and found throughout the world, there is

extremely high potential for an interaction between Avipoxvirus and Myiopsitta monachu.

The interaction between fowl adenovirus HR2 and Netta rufina (red-crested pochard) is also highly possible. Netta rufina belongs to the family Anatidae (waterfowl), which include other ducks, swans, and geese. Fowl adenovirus HR2 is part of the avian adenoviruses (aviadenovirus), which are known to infect waterfowl. Several strains of aviadenovirus have been isolated from different duck species [7] [34].

For interactions between Chicken parvovirus and Poicephalus gulielmi (red-fronted parrot), direct interactions between the two currently appear unlikely. However, Chicken parvovirus belongs to the family Parvoviridae while Poecephalus gulielmi belongs to the family Psittacidae. Interactions between these viruses and hosts from these families are well documented [19]. Furthermore, parvoviruses are often found in areas with large poultry populations [30], heightening the chance of interaction between widespread species such as Poicephalus gulielmi.

Similarly, a 2008 study recorded Avipoxvirus infections in Laysan Albatross in Oahu, Hawaii [37], which suggest that the Avipoxvirus could reasonably also infect Diomedea sanfordi (Northern royal albatross). Additionally, in our interactions data the Diomedea sanfordi has an interaction with Albatrosspox virus, which is in the same family (Poxviridae) as Avipoxvirus.

For the predicted interaction between Melegrivirus A and Gracula religiosa (common hill myna), although Gracula regligiosa has a further distance from known hosts, Melegrivirus A belongs to the Megrivirus genus, which is comprised of multiple viruses that infect a variety of hosts such as duck, pigeon and penguin [23]. This suggests that Melegrivirus A may have high host plasticity and allows it to infect Gracula regligiosa despite the phylogenetic distance.

Edge Score	Virus (taxid)	Host (taxid)	Avg Distance to known Hosts	Avg Distance to Other Species
0.9731	Avipoxvirus isolate PM9 (468380)	Myiopsitta monachus (176066)	159.33	156.40
0.9726	Fowl adenovirus HR2 (911324)	Netta rufina (30387)	161.44	185.58
0.9724	Chicken parvovirus 399/HRV/2010 (1081806)	Poicephalus gulielmi (241588)	207.11	156.79
0.9720	Avipoxvirus isolate Pennsylvania (1294111)	Diomedea sanfordi (2811340)	169.90	171.41
0.9719	Melegrivirus A (1330070)	Gracula religiosa (116992)	207.11	150.39

Table 3: Top 5 most likely predicted previously unobserved virus-host interactions and the host's average phylogenetic distance to known hosts of the virus and average phylogenetic distance to other species.

7 Conclusion and Discussion

This method of utilizing GNNs for link predicting shows promise for being applied to virus-host interactions for both mammalian and avian datasets. This problem is important to detect possible areas of disease spillover and inform surveillance efforts. This has applicability to humans, as they are mammals. An opportunity for future improvement is combining the mammalian and avian datasets into one model, allowing for predicting spillover of avian viruses to mammals and mammalian viruses to birds. This would allow for predicting avian viruses to humans.

Another opportunity for improvement is incorporating a pairwise measure of similarity, such as phylogenetic distance or genome similarity, between host species and virus species into the graph through weighted host-host edges and virus-virus edges. This could be beneficial because viruses may be more likely to infect species that are similar to each other, and likewise hosts might be more susceptible to viruses that are similar.

Additionally, utilizing more robust geographical features of the hosts' range could improve predictions of virus-host interactions. A potential method, as demonstrated by Wardeh et al. [35], intersects host presence maps with geo-layers of features relevant to virus infections, such as climate, agriculture, and human population data. This approach could provide more specific geographical information for each host.

Furthermore, our approach is limited by the quality of existing datasets on avian-virus interactions. There may be certain sampling biases that exist within the dataset. For example, certain avian-virus interactions may be over-reported due to heightened interests (ex. viruses that affect the poultry industry), or accessibility to samples (ex. birds in remote regions may be under-sampled). Certain viruses are also only identified at a Family or Order level and lack specific information on traits such as location of replication or release mechanisms. These missing features add a significant amount of noise to the dataset and lower the accuracy of predictions. Future work could attempt to create a more robust dataset or deal with missing features using strategies such as data imputation.

References

[1] Sonia Altizer, Rebecca Bartel, and Barbara A. Han. 2011. Animal Migration and Infectious Disease Risk. *Science* 331, 6015 (2011), 296–302. https://doi.org/10.1126/science.1194694 arXiv:https://www.science.org/doi/pdf/10.1126/science.1194694

[2] Javed A. Aslam, Emine Yilmaz, and Virgiliu Pavlu. 2005. A geometric interpretation of r-precision and its correlation with average precision. In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (Salvador, Brazil) *(SIGIR '05)*. Association for Computing Machinery, New York, NY, USA, 573–574. https://doi.org/10.1145/1076034.1076134

[3] Bethesda (MD): National Center for Biotechnology Information (US). 2010. Entrez Programming Utilities Help [Internet]. https://www.ncbi.nlm.nih.gov/books/NBK25501/

[4] Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information. 2004. NCBI Virus [Internet]. https://www.ncbi.nlm.nih.gov/labs/virus/vssi/#/

[5] TR Boosinger, RW Winterfield, DS Feldman, and AS Dhillon. 1982. Psittacine pox virus: virus isolation and identification, transmission, and cross-challenge

studies in parrots and chickens. *Avian diseases* (1982), 437–444.
[6] Dennis Carroll, Peter Daszak, Nathan D Wolfe, George F Gao, Carlos M Morel, Subhash Morzaria, Ariel Pablos-Méndez, Oyewale Tomori, and Jonna AK Mazet. 2018. The global virome project. *Science* 359, 6378 (2018), 872–874.
[7] Shilong Chen, Fengqiang Lin, Bin Jiang, Shifeng Xiao, Dandan Jiang, Chang Lin, Shao Wang, Xiaoxia Cheng, Xiaoli Zhu, Hui Dong, et al. 2022. Isolation and characterization of a novel strain of duck aviadenovirus B from Muscovy ducklings with acute hepatitis in China. *Transboundary and Emerging Diseases* 69, 5 (2022), 2769–2778.
[8] NCBI Resource Coordinators. 2018. Database resources of the national center for biotechnology information. *Nucleic acids research* 46, Database issue (2018), D8.
[9] Zhi-Hua Du, Jun-Peng Zhong, Yun Liu, and Jian-Qiang Li. 2023. Prokaryotic virus host prediction with graph contrastive augmentaion. *PLOS Computational Biology* 19, 12 (2023), e1011671.
[10] Felipe CB Esteves, Sandra Y Marín, Maurício Resende, Aila SG Silva, Hannah LG Coelho, Mayara B Barbosa, Natália S D'Aparecida, José S de Resende, Ana CD Torres, and Nelson RS Martins. 2017. Avian pox in native captive Psittacines, Brazil, 2015. *Emerging Infectious Diseases* 23, 1 (2017), 154.
[11] Matthias Fey and Jan Eric Lenssen. 2019. Fast Graph Representation Learning with PyTorch Geometric. arXiv:1903.02428 [cs.LG] https://arxiv.org/abs/1903.02428
[12] William L. Hamilton, Rex Ying, and Jure Leskovec. 2018. Inductive Representation Learning on Large Graphs. arXiv:1706.02216 [cs.SI] https://arxiv.org/abs/1706.02216
[13] Chantal Hulo, Edouard De Castro, Patrick Masson, Lydie Bougueleret, Amos Bairoch, Ioannis Xenarios, and Philippe Le Mercier. 2011. ViralZone: a knowledge resource to understand virus diversity. *Nucleic acids research* 39, suppl_1 (2011), D576–D582.
[14] RM Imrie, KE Roberts, and B Longdon. 2021. Between virus correlations in the outcome of infection across host species: Evidence of virus by host species interactions. *Evol Lett* 5, 5 (2021), 472–483. https://doi.org/10.1002/evl3.247
[15] Walter Jetz, Gavin H. Thomas, Jeffrey B. Joy, Klaas Hartmann, and Arne O. Mooers. 2012. The global diversity of birds in space and time. *Nature* 491 (2012), 444–448.
[16] Kate E Jones, Nikkita G Patel, Marc A Levy, Adam Storeygard, Deborah Balk, John L Gittleman, and Peter Daszak. 2008. Global trends in emerging infectious diseases. *Nature* 451, 7181 (2008), 990–993.
[17] Diederik P. Kingma and Jimmy Ba. 2017. Adam: A Method for Stochastic Optimization. arXiv:1412.6980 [cs.LG] https://arxiv.org/abs/1412.6980
[18] Michael Klear. 2019. dfencoder. https://github.com/AlliedToasters/dfencoder. Accessed: 2024-12-02.
[19] Natalie Klukowski, Paul Eden, Muhammad Jasim Uddin, and Subir Sarker. 2024. Virome of Australia's most endangered parrot in captivity evidenced of harboring hitherto unknown viruses. *Microbiology spectrum* 12, 1 (2024), e03052–23.
[20] Mikaela Koutrouli, Evangelos Karatzas, David Paez-Espino, and Georgios A Pavlopoulos. 2020. A guide to conquer the biological network era using graph theory. *Frontiers in bioengineering and biotechnology* 8 (2020), 34.
[21] Christine Kreuder Johnson, Peta L Hitchens, Tierra Smiley Evans, Tracey Goldstein, Kate Thomas, Andrew Clements, Damien O Joly, Nathan D Wolfe, Peter Daszak, William B Karesh, et al. 2015. Spillover and pandemic properties of zoonotic viruses with high host plasticity. *Scientific reports* 5, 1 (2015), 14830.
[22] Guillaume Lachaud, Patricia Conde-Cespedes, and Maria Trocan. 2022. Comparison between Inductive and Transductive Learning in a Real Citation Network using Graph Neural Networks. In *2022 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. 534–540. https://doi.org/10.1109/ASONAM55673.2022.10068589
[23] Elliot J. Lefkowitz, Donna M. Dempsey, Robert C. Hendrickson, Richard J. Orton, Stuart G. Siddell, and David B. Smith. 2018. Virus taxonomy: The database of the International Committee on Taxonomy of Viruses (ICTV). *Nucleic Acids Research* 46 (2018), D708–D717. https://doi.org/10.1093/nar/gkx932
[24] Denis Lepage, Gaurav Vaidya, and Robert Guralnick. 2014. Avibase – a database system for managing and organizing taxonomic concepts. *ZooKeys* 420 (Jun 2014), 117–135. https://doi.org/10.3897/zookeys.420.7089
[25] Ryan Lichtnwalter and Nitesh V. Chawla. 2012. Link Prediction: Fair and Effective Evaluation. In *2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*. 376–383. https://doi.org/10.1109/ASONAM.2012.68
[26] Kevin J Olival, Parviez R Hosseini, Carlos Zambrana-Torrelio, Noam Ross, Tiffany L Bogich, and Peter Daszak. 2017. Host and viral traits predict zoonotic spillover from mammals. *Nature* 546, 7660 (2017), 646–650.
[27] World Health Organization. [n. d.]. WHO EMRO: | Zoonotic disease: emerging public health threats in the Region. https://www.emro.who.int/about-who/rc61/zoonotic-diseases.html/
[28] Timothée Poisot, M. Annie Ouellet, Nicolas Mollentze, Matthew J. Farrell, Daniel J. Becker, Leonard Brierley, Gregory F. Albery, Rory J. Gibb, Stephanie N. Seifert, and Colin J. Carlson. 2023. Network embedding unveils the hidden interactions in the mammalian virome. *Patterns (N Y)* 4, 6 (2023), 100738. https://doi.org/10.1016/j.patter.2023.100738
[29] Kurt D Reed, Jennifer K Meece, James S Henkel, and Sanjay K Shukla. 2003. Birds, migration and emerging zoonoses: West Nile virus, Lyme disease, influenza A and enteropathogens. *Clinical medicine & research* 1, 1 (2003), 5–12.
[30] Christian Sánchez, Ana Doménech, Esperanza Gomez-Lucia, José Luis Méndez, Juan Carlos Ortiz, and Laura Benítez. 2023. A Novel Dependoparvovirus Identified in Cloacal Swabs of Monk Parakeet (Myiopsitta monachus) from Urban Areas of Spain. *Viruses* 15, 4 (2023), 850.
[31] Jiayu Shang and Yanni Sun. 2021. Predicting the hosts of prokaryotic viruses using GCN-based semi-supervised learning. *BMC biology* 19 (2021), 1–15.
[32] Jiayu Shang and Yanni Sun. 2022. CHERRY: a Computational metHod for accuratE pReDiction of virus–pRokarYotic interactions using a graph encoder–decoder model. *Briefings in Bioinformatics* 23, 5 (September 2022), bbac182. https://doi.org/10.1093/bib/bbac182
[33] Joseph A. Tobias, Catherine Sheard, Alex L. Pigot, Adam J. M. Devenish, Jingyi Yang, Ferran Sayol, Montague H. C. Neate-Clegg, Nico Alioravainen, Thomas L. Weeks, Robert A. Barber, Patrick A. Walkden, Hannah E. A. MacGregor, Samuel E. I. Jones, Claire Vincent, Anna G. Phillips, Nicola M. Marples, Flavia A. Montaño-Centellas, Victor Leandro-Silva, Santiago Claramunt, Bianca Darski, Benjamin G. Freeman, Tom P. Bregman, Christopher R. Cooney, Emma C. Hughes, Elliot J. R. Capp, Zoë K. Varley, Nicholas R. Friedman, Heiko Korntheuer, Andrea Corrales-Vargas, Christopher H. Trisos, Brian C. Weeks, Dagmar M. Hanz, Till Töpfer, Gustavo A. Bravo, Vladimír Remeš, Larissa Nowak, Lincoln S. Carneiro, Amilkar J. Moncada R., Beata Matysioková, Daniel T. Baldassarre, Alejandra Martínez-Salinas, Jared D. Wolfe, Philip M. Chapman, Benjamin G. Daly, Marjorie C. Sorensen, Alexander Neu, Michael A. Ford, Rebekah J. Mayhew, Luis Fabio Silveira, David J. Kelly, Nathaniel N. D. Annorbah, Henry S. Pollock, Ada M. Grabowska-Zhang, Jay P. McEntee, Juan Carlos T. Gonzalez, Camila G. Meneses, Marcia C. Muñoz, Luke L. Powell, Gabriel A. Jamie, Thomas J. Matthews, Oscar Johnson, Guilherme R. R. Brito, Kristof Zyskowski, Ross Crates, Michael G. Harvey, Maura Jurado Zevallos, Peter A. Hosner, Tom Bradfer-Lawrence, James M. Maley, F. Gary Stiles, Hevana S. Lima, Kaiya L. Provost, Moses Chibesa, Mmatjie Mashao, Jeffrey T. Howard, Edson Mlamba, Marcus A. H. Chua, Bicheng Li, M. Isabel Gómez, Natalia C. García, Martin Päckert, Jérôme Fuchs, Jarome R. Ali, Elizabeth P. Derryberry, Monica L. Carlson, Rolly C. Urriza, Kristin E. Brzeski, Dewi M. Prawiradilaga, Matt J. Rayner, Eliot T. Miller, Rauri C. K. Bowie, René-Marie Lafontaine, R. Paul Scofield, Yingqiang Lou, Lankani Somarathna, Denis Lepage, Marshall Illif, Eike Lena Neuschulz, Mathias Templin, D. Matthias Dehling, Jacob C. Cooper, Olivier S. G. Pauwels, Kangkuso Analuddin, Jon Fjeldså, Nathalie Seddon, Paul R. Sweet, Fabrice A. J. DeClerck, Luciano N. Naka, Jeffrey D. Brawn, Alexandre Aleixo, Katrin Böhning-Gaese, Carsten Rahbek, Susanne A. Fritz, Gavin H. Thomas, and Matthias Schleuning. 2022. AVONET: morphological, ecological and geographical data for all birds. *Ecology Letters* 25, 3 (2022), 581–597. https://doi.org/10.1111/ele.13898 arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/ele.13898
[34] Jessy Vibin, Anthony Chamings, Marcel Klaassen, and Soren Alexandersen. 2020. Metagenomic characterisation of additional and novel avian viruses from Australian wild ducks. *Scientific Reports* 10, 1 (2020), 22284.
[35] Mohamed Wardeh, Mark S. C. Blagrove, Kieran J. Sharkey, Matthew Baylis, and James L. N. Wood. 2021. Divide-and-conquer: machine-learning integrates mammalian and viral traits with network features to predict virus-mammal associations. *Nature Communications* 12 (2021), 3954. https://doi.org/10.1038/s41467-021-24085-w
[36] Zonghan Wu, Shirui Pan, Fengwen Chen, Guodong Long, Chengqi Zhang, and Philip S. Yu. 2021. A Comprehensive Survey on Graph Neural Networks. *IEEE Transactions on Neural Networks and Learning Systems* 32, 1 (2021), 4–24. https://doi.org/10.1109/TNNLS.2020.2978386
[37] Lindsay Young and Eric VanderWerf. 2008. Prevalence of avian pox virus and effect on the fledging success of Laysan Albatross. *Journal of Field Ornithology* 79 (03 2008), 93 – 98. https://doi.org/10.1111/j.1557-9263.2008.00149.x

Predicting Missing Virus-Host Links for Avian Hosts using Graph Neural Networks

Akshath Shvetang Anna
aanna7@gatech.edu
Georgia Institute of Technology
Atlanta, Georgia, USA

Grace Driskill
gdriskill3@gatech.edu
Georgia Institute of Technology
Atlanta, Georgia, USA

Julia Qian
jqian310@gatech.edu
Georgia Institute of Technology
Atlanta, Georgia, USA

Abstract

A virus's host range is an important indicator of its zoonotic potential. Predicting the missing hosts of viruses is thus an important step in supporting zoonotic disease surveillance, outbreak prevention and outbreak response. Here, we survey current methods and approaches in predicting missing links between animal species and pathogens. We then propose a graph neural network model to predict associations between bird species and viruses, using an aggregated dataset with bird and virus features.

ACM Reference Format:
Akshath Shvetang Anna, Grace Driskill, and Julia Qian. 2018. Predicting Missing Virus-Host Links for Avian Hosts using Graph Neural Networks. In *CSE 8803 EPI, Fall, 2024, Atlanta, GA.* ACM, New York, NY, USA, 5 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

A zoonotic disease is an infectious illness transmitted from animals to humans. Approximately 60% of all human illnesses originate from animals [10], including high-profile diseases such as Ebola and COVID-19 (both believed to originate from bats). The WHO estimates that approximately 1 billion cases of zoonoses occur annually, with millions of deaths [16]. Avian zoonoses, originating from birds, present a unique challenge regarding surveillance and predictions. Birds, like mammals, can act as hosts to viruses. Alternatively, they can spread the disease to vectors which then pass the infection to humans (such as the bird-mosquito-human pathway for West Nile virus). Moreover, birds can greatly increase the geographic range of a disease by spreading it anywhere along their migration routes [19], infecting both humans and domesticated animals.

Spillover is the event in which a disease isolated to one species begins infecting another. The range of species a virus can infect (host plasticity) is directly related to its potential to spillover to humans and its geographic reach [12]. Predicting a virus's host range can grant insight into which disease may become zoonotic, helping forecast potential outbreaks and supporting surveillance efforts.

CSE 8803 EPI, November 2024, Atlanta, GA, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM
https://doi.org/XXXXXXX.XXXXXXX

Research efforts in this area by Wardeh et al. [23] have leveraged machine learning to predict previously unknown connections between mammal species and known viruses. Their results reveal over 20,000 potential unknown associations between viruses and mammals, pointing to how small a fraction of virus-mammal associations are known to researchers. Although their model provides extensive predictions for mammalian hosts, with 1.67 million undiscovered viruses that exist in mammals and bird hosts [4], this remains an incomplete picture.

Host-virus interactions, as well as ecological interactions as a whole, can be represented through biological networks [11]. Increasingly, machine learning techniques, in particular neural networks, have been applied to such networks as a means to analyze and identify patterns within data. Common tasks include node or edge classification, node clustering, link prediction, and recommendation [26]. Within the area of zoonotic diseases, graph neural networks (GNNs) have been leveraged to predict viral hosts, typically for prokaryotic viruses [20]. Virus-host predictions can be interpreted as a link prediction problem, where predictions are made about the probability of an edge forming between a host and a virus node.

This project aims to apply GNNs to analyze host-pathogen and interspecies avian datasets and predict unknown associations between birds and viruses. This will grant better insight into a virus's range of hosts and by extension their potential to spill over to humans.

2 Addressing Proposal Comments

We have given a more descriptive name to the project.

Additionally, we have enlarged the scope of our project from our proposal. We initially proposed to replicate the approach used in Wardeh et al. [23], focusing on generating a dataset with avian-virus interactions. Instead, we have expanded our scope with the novel application of GNNs to predict potential unknown avian virus hosts. We will leverage Wardeh et al. [23]'s dataset and results as a starting point to evaluate the performance of our approach. We will then apply our model to a new dataset of avian-virus interactions with limited features to predict unknown avian virus hosts.

Finally, while our approach cannot be directly used for outbreak detection, gaining a clearer understanding of what avian species (and hosts in general) can host certain viruses is the first step in anticipating and forecasting potential novel outbreaks. For example, if a specific genus of waterfowl is predicted to host a highly transmissible influenza with a large host range, policymakers could provide both warnings to individuals in high contact with these birds and recommend preventative measures to avoid spillover into humans. These applications will be further discussed in the final report's Discussion section.

3 Survey of Related Work

Given the global health burden presented by zoonotic disease and the relationship between a virus's host range and its zoonotic potential [12], there has been an increasing interest in understanding the full scope of a virus's existing and potential hosts.

Wardeh et al. [23] aimed to predict unknown associations between mammals and viruses through ensemble learning. They created three machine learning models to predict virus-mammal associations, mammal-virus associations, and missing edges in a global bipartite network, before applying ensemble voting. Although their paper is focused on mammal-virus associations, their goal to predict unknown animal-virus associations provides insight into how an assessment of avian-virus associations can be approached. One limitation of their approach is that, by using interaction data, their results may be biased due to targeted research. Associations may be over or under-reported due to factors like geographical accessibility (i.e. it may be harder to reach some populations of animals for testing), or an increased interest in certain mammals due to current outbreaks. Their global network view provides an alternative method to deal with these research and sampling biases.

Olival et al. [14] investigate factors that impact a virus's host range and transmissibility to other species. They discover that viral richness, the number of different viruses found in a host species, is impacted by research effort, mammal sympatry (when two species live in the same area) and host geographic range. In turn, factors that impact a mammal species's zoonotic viral richness include overall viral richness, phylogenetic distance from humans, host order and research effort. Although their research targets human-virus associations, their research highlights important geographical and phylogenetic factors that impact a virus's transmission from one species to another. We can apply the features they found that impact viral richness to our model to better estimate how many viruses different avian-species host. Similar to the previous two papers, their work is also potentially biased by the attention each virus receives in the research space and cannot account for associations that currently remain undiscovered.

Poisot et al. [18] address the challenge of predicting missing host-virus interactions as a network problem. They represent the network as a bipartite structure, where virus nodes connect to host nodes where an association is known. They used a combined linear filtering and singular value decomposition model with hyper-parameters for determining the importance of different network structure characteristics. By choosing hyper-parameters that emphasize the network's connectance instead of in-degree or out-degree, they helped to reduce the bias caused by uneven research efforts, where certain parts of the network have more data due to concentrated study. Their techniques lead to predictions for possible new host-virus interactions that are robust to these data biases and suggest that future efforts for virus discovery be focused on the Amazon Basin and sub-Saharan Africa. While their method effectively handled the issue of incomplete and unbalanced data, a limitation was the exclusion of host and viral traits as predictors, which could enhance the model's performance.

Altizer et al. [1] investigates the relationship between animal migration and zoonotic diseases, suggesting that migration can both facilitate and mitigate disease spreading. A common concern is that migratory species may move diseases long distances geographically, and although this is a possibility, there are few documented examples of this phenomenon. More likely concerns are that the migratory species themselves are exposed to a wider range of diseases throughout their migration and that hotspots for diseases are created in places where multiple species of migratory animals stop to rest. An example of this is Delaware Bay, where stopping shorebirds create a hotspot of avian influenza. The stress of migration may lower the animals' immune system making them more susceptible to disease. However, Altizer et al. [1] also suggest that migration can actually decrease the spread of disease because staying in one habitat for a long period of time allows parasites to accumulate. Additionally, migration removes diseased members of a species from a community, and then decreases the chances of them spreading the illness, because they are not able to complete the migration.

This information is important when considering what factors to take into account when trying to predict potential zoonotic diseases. Due to the complicated role that migration has in zoonotic disease transmission, including migratory animals and information on their migrations could impact the model.

Shang and Sun [21] studied a similar problem of predicting hosts of prokaryotic viruses, which is relevant in the microbiology field. They approached host prediction as a link prediction task in a knowledge graph created from protein and DNA sequence features of the involved viruses and hosts, such as protein organization and sequence similarity. Their model has an encoder-decoder structure with a graph convolution network (GCN) encoder to utilize the topological structure and embedded node features from the graph. The decoder estimates the probability of a virus-prokaryote pair having a link. This study reports a sizable improvement in host prediction accuracy by using this model, from 43% to 80%. This suggests that GNNs and similar graph-based architecture could be promising for predicting virus-host interactions for zoonotic disease too. Additionally, this study's model only uses sequence data without integrating other traits of the hosts or viruses, which is a potential place for improvement.

Similarly, Du et al. [6] applied GNNs to predict phage-host interactions, with the intent to overcome the constraints that exist on current prediction models caused by the number of potential unknown interactions in comparison to the datasets available. They construct a heterogeneous phage-prokaryote graph using virus-virus protein and virus-host DNA similarities. They then leveraged a light graph convolutional network (LGCN) to recommend the host with the highest probability for each virus and achieved results competitive with existing virus-host interaction prediction methods. However, their constructed network did not include host-host interactions, which could improve the quality and accuracy of their predictions.

4 Problem Formulation

The problem we aim to address is that, while some associations between hosts and viruses are known, many associations between hosts and known viruses remain unknown. Using a dataset of bird traits, a dataset of virus traits and a dataset of bird-viruses associations, our goal is to extract key features from each to predict these missing associations between birds and viruses.

We will treat this as a link-prediction problem and apply GNNs to predict links between viral and avian nodes. We will first apply our GNN model to the Divide-and-conquer dataset of mammal-virus interactions [23]. This will allow us to evaluate the performance of our approach against the predictions made by Wardeh et al. [23]'s three-perspective approach using a robust dataset containing phylogenetic, ecological, and genetic traits. Then, we will apply our GNN model to our own constructed dataset of avian-virus interactions containing a smaller feature set. In other words, we will predict which birds are potential hosts to known viruses, with an emphasis on predicting previously unreported relationships.

5 Data

For this project, we plan to use the datasets used in Wardeh et al. [23] for information on mammal-virus interactions, mammal traits and mammalian viruses. Additionally, we are creating additional datasets on virus-avian interactions, avian traits and avian viruses.

5.1 Virus-Avian Interactions Dataset

We initially planned to obtain virus-avian interaction from ENHanCEd Infectious Diseases Database (EID2)[24], the database used in Wardeh et al. [23] for virus-mammal associations, along with the Bacterial and Viral Bioinformatics Resource Center (BV-BRC) [15]. However, we were unable to contact the EID2 team to access their data and the host names from BV-BRC were inconsistent, so we had to switch directions from our original purposed plan.

Virus-bird interactions were downloaded from the NCBI Virus Database [3] by filtering the Virus/Taxonomy for Viruses, taxid=10239 and Host for Aves (birds), taxid=8782. 347,583 records were downloaded in a CSV format. The cleaned CSV files contains only the following columns:

- Pathogen Species - Species in the original data
- Host - scientific name of the avian host that the pathogen infected
- Total Count - total number of entries in the original data for this pathogen, host pair
- GenBank - number of entries in the original data that are from GenBank for this pathogen, host pair
- RefSeq - number of entries in the original data that are from RefSeq for this pathogen, host pair
- Completed Count - number of entries in the original data with Nuc_Completeness = complete for this pathogen, host pair

From the NCBI Virus Database, 4743 virus-bird interactions were found. This included 1184 unique avian hosts and 1929 unique viruses.

5.2 Avian Dataset

A dataset of the avian hosts identified in the virus-host interactions and their traits was constructed. The virus Taxon ID and species lineage (order, family, and genus) are obtained from NCBI Taxonomy Database using NCBI's web API Entrez Programming Utilities [2].

The traits for each bird includes phylogenetic, ecological and geospatial traits, mirroring the mammal traits used in Wardeh et al. [23]. The following sources are used for this: The EltonTraits 1.0 [25] dataset for information on birds' diet and body mass, AVONET [22] for information on birds' geographical range, migration, trophic level and niche, and primary lifestyle, and BirdTree [9] for birds' phylogeny and taxonomy.

5.3 Virus Dataset

We also built a dataset for avian viruses and their traits. For each virus species in the virus-avian interaction dataset, the official virus species name is taken from the NCBI Virus Database [3] and ICTV [13]. The virus Taxon ID and species lineage (order, family, and genus) are obtained from the NCBI Taxonomy Database using NCBI's web API Entrez Programming Utilities [2].

The viral genome composition is obtained from ViralZone [8] and ICTV. These include the viral genome composition (1=RNA, 0=DNA), retrotranscription (1=yes, 0=no), if the virus has double-stranded genetic material (1=double stranded, 0=single stranded), cytoplasmic replication (1=yes, 0=no), if the virus is enveloped (1=yes, 0=no), if the genome is circular (1=circular, 0=linear), if the genetic material is positive or negative sense, and if the virus is segmented or monopartite.

Genome-specific information is also included. Genomic size (in KB), GC content, and number of coding segments are taken from the NCBI Genome Assembly database [5].

Both ViralZone and the ICTV database are used to get information on replication sites, viral release mechanisms, and methods of cell entry.

Traits are assigned for each virus species. Species without data on certain traits are assigned traits one level up their lineage (i.e. genus, then order, then family).

6 Approach

Wardeh et al. [23] leveraged a unique approach with majority voting between models of the following three perspectives:

(1) a viral perspective that explored which potential mammalian hosts each virus could infect, based on the host characteristics
(2) a mammalian perspective that explored which viruses could potentially infect each mammal based on the viral characteristics
(3) a network perspective that is organized as a bipartite graph with viral and mammalian sets

This approach allowed them to balance potential biases due to the available information on viruses, hosts, and known connections. Two of the perspectives were so-called "local" perspectives that considered each virus and mammalian host, and the final perspective was a global understanding of virus-mammal associations. Indeed, the majority voting between these perspectives led to better results than any one or combination of them. Despite these benefits, there is added complexity in training numerous classifiers across each of these perspectives, with the local perspectives requiring different classifiers for each individual virus in the virus perspective and each mammal in the mammalian perspective.

Graph Neural Networks present an opportunity to incorporate information about all of these aspects into a singular model. Without the separation of perspectives and majority voting, the results

may not be as robust, however the benefit would be greater simplicity.

We will be utilizing a bipartite graph. The virus pathogens will be one class of nodes, and the hosts will be the other class (starting with mammalian hosts and then moving on to avian hosts). The edges will represent virus-host associations. GNNs can be used for predictions tasks at the node, edge, and global level. The target task of this study, link-prediction is an edge-level task [27].

The information collected on these entities will be incorporated into node features. Numerous approaches exist for training the GNNs. Özer et al. [17] analyzed numerous methodologies from non-learning link scoring to learning-based embedding-based link prediction. Many available approaches were not developed for bipartite graphs and are not great matches for this problem specification. GNNs have long been used for recommendation tasks, and interestingly, they found that embedding-based methods for these recommendation tasks can be effectively modified for general link-prediction tasks on bipartite networks. They noted that the performance greatly varied on the dataset being observed, so we will assess the two top overall performers among these recommender-based approaches: DiffRec and NGCF.

7 Evaluation

Inherently, there is a class-imbalance between positive instances (edges) and negative instances (no edge). The Area Under the Receiver Operating Characteristic Curve (AUC) is not ideal in these cases, and the Area Under the Precision-Recall Curve (AUPR) is better, so we will use that instead [17].

8 Remaining Work

8.1 Data

In our proposal, we intended to use the Enhanced Infectious Diseases (EID2) database [24] to compile our database of avian-virus interactions. However, we still have not heard back from EID2. If we receive a reply from their team, we will integrate their data into our current dataset. However, EID2 pulls data from the NCBI Taxonomy database and the NCBI Nucleotide database, similar to the sources of the NCBI Virus Database. Thus, we anticipate a large overlap between the two databases.

While we currently aim to create a database outlined with the features discussed in Section 5, our dataset can be extended to include phylogenetic data. We will investigate if there is a scalable and consistent method of calculating phylogenetic distance and host breadth for viral and host genomes.

We are still working on completing the viral dataset. One issue is inconsistency with naming and missing information on certain species of viruses. We aim to resolve this by assigning viruses traits a level up their taxonomy, or through manual assignment.

Similarly, we are still working on the avian dataset. Some of the hosts from the virus-avian interactions dataset are not identified at a species level, but instead at a higher taxonomic level. We will resolve this by assigning these hosts traits for the taxonomic level they are identified at.

8.2 Approach

The first step in our process is to compare performance of GNN link prediction to the results of Wardeh et al. [23]. To that end, we needed to collect the dataset used by them for virus-mammal predictions. The files they uploaded included un-merged datasets. There were included scripts for generating the merged datasets, and all of the code was written in R. Thus, effort was spent on learning the basics of programming in R so that the source code could be correctly interpreted. Additional delays were encountered due to missing lines and typos in the source code. The data has been generated now, with information of 556 different viruses and 698 mammalian hosts.

Due to change in our project direction, we have not yet completed the implementation of our GNN training and evaluation scripts, and this will be the main priority going forward.

8.3 Evaluation

Once the GNN scripts are completed, we will predict missing links for virus-mammal interactions and compare the results against those of Wardeh et al. [23]. We will choose an encoding method and make further modifications based on what makes the predicted results most comparable.

After this, we will apply the same methodology to the avian dataset that we have constructed to predict missing links between the expanded set of viruses we are exploring and the avian hosts. We will run these experiments multiple times and provide confidence intervals.

8.4 Discussion

We will discuss the applications of our approach in practice, including its potential applications and limits in outbreak detection and response (see Section 2).

8.5 Future Directions

Here, we will discuss any limitations of our approach and areas for future improvement. One important limitation of our approach is the size of our avian-virus interactions dataset. We did not include phylogenetic or evolutionary distance features, which have been included traditionally in interaction-prediction approaches [7]. This can be addressed by adding more viral and avian features to the dataset, such as host breadth (for viruses) or evolutionary distance between viral and avian genomes.

References

[1] Sonia Altizer, Rebecca Bartel, and Barbara A. Han. 2011. Animal Migration and Infectious Disease Risk. *Science* 331, 6015 (2011), 296–302. https://doi.org/10.1126/science.1194694 arXiv:https://www.science.org/doi/pdf/10.1126/science.1194694

[2] Bethesda (MD): National Center for Biotechnology Information (US). 2010. Entrez Programming Utilities Help [Internet]. https://www.ncbi.nlm.nih.gov/books/NBK25501/

[3] Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information. 2004. NCBI Virus [Internet]. https://www.ncbi.nlm.nih.gov/labs/virus/vssi/#/

[4] Dennis Carroll, Peter Daszak, Nathan D Wolfe, George F Gao, Carlos M Morel, Subhash Morzaria, Ariel Pablos-Méndez, Oyewale Tomori, and Jonna AK Mazet. 2018. The global virome project. *Science* 359, 6378 (2018), 872–874.

[5] NCBI Resource Coordinators. 2018. Database resources of the national center for biotechnology information. *Nucleic acids research* 46, Database issue (2018), D8.

[6] Zhi-Hua Du, Jun-Peng Zhong, Yun Liu, and Jian-Qiang Li. 2023. Prokaryotic virus host prediction with graph contrastive augmentaion. *PLOS Computational Biology* 19, 12 (2023), e1011671.

[7] Maxwell J Farrell, Mohamad Elmasri, David A Stephens, and T Jonathan Davies. 2022. Predicting missing links in global host–parasite networks. *Journal of Animal Ecology* 91, 4 (2022), 715–726.

[8] Chantal Hulo, Edouard De Castro, Patrick Masson, Lydie Bougueleret, Amos Bairoch, Ioannis Xenarios, and Philippe Le Mercier. 2011. ViralZone: a knowledge resource to understand virus diversity. *Nucleic acids research* 39, suppl_1 (2011), D576–D582.

[9] Walter Jetz, Gavin H. Thomas, Jeffrey B. Joy, Klaas Hartmann, and Arne O. Mooers. 2012. The global diversity of birds in space and time. *Nature* 491 (2012), 444–448.

[10] Kate E Jones, Nikkita G Patel, Marc A Levy, Adam Storeygard, Deborah Balk, John L Gittleman, and Peter Daszak. 2008. Global trends in emerging infectious diseases. *Nature* 451, 7181 (2008), 990–993.

[11] Mikaela Koutrouli, Evangelos Karatzas, David Paez-Espino, and Georgios A Pavlopoulos. 2020. A guide to conquer the biological network era using graph theory. *Frontiers in bioengineering and biotechnology* 8 (2020), 34.

[12] Christine Kreuder Johnson, Peta L Hitchens, Tierra Smiley Evans, Tracey Goldstein, Kate Thomas, Andrew Clements, Damien O Joly, Nathan D Wolfe, Peter Daszak, William B Karesh, et al. 2015. Spillover and pandemic properties of zoonotic viruses with high host plasticity. *Scientific reports* 5, 1 (2015), 14830.

[13] Elliot J. Lefkowitz, Donna M. Dempsey, Robert C. Hendrickson, Richard J. Orton, Stuart G. Siddell, and David B. Smith. 2018. Virus taxonomy: The database of the International Committee on Taxonomy of Viruses (ICTV). *Nucleic Acids Research* 46 (2018), D708–D717. https://doi.org/10.1093/nar/gkx932

[14] Kevin J Olival, Parviez R Hosseini, Carlos Zambrana-Torrelio, Noam Ross, Tiffany L Bogich, and Peter Daszak. 2017. Host and viral traits predict zoonotic spillover from mammals. *Nature* 546, 7660 (2017), 646–650.

[15] Robert D. Olson, Rima Assaf, Thomas Brettin, Neil Conrad, Cody Cucinell, James J. Davis, Donna M. Dempsey, Alan Dickerman, Elizabeth M. Dietrich, Robert W. Kenyon, Melis Kuscuoglu, Elliot J. Lefkowitz, James Lu, Danielle Machi, Catherine Macken, Chunhong Mao, Anna Niewiadomska, Maria Nguyen, Gary J. Olsen, Ross C. Overbeek, Bryan Parrello, Vincent Parrello, Joshua S. Porter, Gordon D. Pusch, Maulik Shukla, Indu Singh, Leon Stewart, Grace Tan, Cathy Thomas, Mary VanOeffelen, Veronika Vonstein, Zachary S. Wallace, Alexander S. Warren, Alice R. Wattam, Fangfang Xia, Haechan Yoo, Yan Zhang, Christian M. Zmasek, Richard H. Scheuermann, and Rick L. Stevens. 2023. Introducing the Bacterial and Viral Bioinformatics Resource Center (BV-BRC): a resource combining PATRIC, IRD and ViPR. *Nucleic Acids Research* 51, D1 (2023), D678–D689. https://doi.org/10.1093/nar/gkac1003

[16] World Health Organization. [n. d.]. WHO EMRO: | Zoonotic disease: emerging public health threats in the Region. https://www.emro.who.int/about-who/rc61/zoonotic-diseases.html/

[17] Şükrü Demir Inan Özer, Günce Keziban Orman, and Vincent Labatut. 2024. Link Prediction in Bipartite Networks. In *28th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems (KES)*. Sevilla, Spain. https://hal.science/hal-04605758

[18] Timothée Poisot, M. Annie Ouellet, Nicolas Mollentze, Matthew J. Farrell, Daniel J. Becker, Leonard Brierley, Gregory F. Albery, Rory J. Gibb, Stephanie N. Seifert, and Colin J. Carlson. 2023. Network embedding unveils the hidden interactions in the mammalian virome. *Patterns (N Y)* 4, 6 (2023), 100738. https://doi.org/10.1016/j.patter.2023.100738

[19] Kurt D Reed, Jennifer K Meece, James S Henkel, and Sanjay K Shukla. 2003. Birds, migration and emerging zoonoses: West Nile virus, Lyme disease, influenza A and enteropathogens. *Clinical medicine & research* 1, 1 (2003), 5–12.

[20] Jiayu Shang and Yanni Sun. 2021. Predicting the hosts of prokaryotic viruses using GCN-based semi-supervised learning. *BMC biology* 19 (2021), 1–15.

[21] Jiayu Shang and Yanni Sun. 2022. CHERRY: a Computational metHod for accuratE pRediction of virus–pRokarYotic interactions using a graph encoder–decoder model. *Briefings in Bioinformatics* 23, 5 (September 2022), bbac182. https://doi.org/10.1093/bib/bbac182

[22] Joseph A. Tobias, Catherine Sheard, Alex L. Pigot, Adam J. M. Devenish, Jingyi Yang, Ferran Sayol, Montague H. C. Neate-Clegg, Nico Alioravainen, Thomas L. Weeks, Robert A. Barber, Patrick A. Walkden, Hannah E. A. MacGregor, Samuel E. I. Jones, Claire Vincent, Anna G. Phillips, Nicola M. Marples, Flavia A. Montaño-Centellas, Victor Leandro-Silva, Santiago Claramunt, Bianca Darski, Benjamin G. Freeman, Tom P. Bregman, Christopher R. Cooney, Emma C. Hughes, Elliot J. R. Capp, Zoë K. Varley, Nicholas R. Friedman, Heiko Korntheuer, Andrea Corrales-Vargas, Christopher H. Trisos, Brian C. Weeks, Dagmar M. Hanz, Till Töpfer, Gustavo A. Bravo, Vladimír Remeš, Larissa Nowak, Lincoln S. Carneiro, Amilkar J. Moncada R., Beata Matysioková, Daniel T. Baldassarre, Alejandra Martínez-Salinas, Jared D. Wolfe, Philip M. Chapman, Benjamin G. Daly, Marjorie C. Sorensen, Alexander Neu, Michael A. Ford, Rebekah J. Mayhew, Luis Fabio Silveira, David J. Kelly, Nathaniel N. D. Annorbah, Henry S. Pollock, Ada M. Grabowska-Zhang, Jay P. McEntee, Juan Carlos T. Gonzalez, Camila G. Meneses, Marcia C. Muñoz, Luke L. Powell, Gabriel A. Jamie, Thomas J. Matthews, Oscar Johnson, Guilherme R. R. Brito, Kristof Zyskowski, Ross Crates, Michael G. Harvey, Maura Jurado Zevallos, Peter A. Hosner, Tom Bradfer-Lawrence, James M. Maley, F. Gary Stiles, Hevana S. Lima, Kaiya L. Provost, Moses Chibesa, Mmatjie Mashao, Jeffrey T. Howard, Edson Mlamba, Marcus A. H. Chua, Bicheng Li, M. Isabel Gómez, Natalia C. García, Martin Päckert, Jérôme Fuchs, Jarome R. Ali, Elizabeth P. Derryberry, Monica L. Carlson, Rolly C. Urriza, Kristin E. Brzeski, Dewi M. Prawiradilaga, Matt J. Rayner, Eliot T. Miller, Rauri C. K. Bowie, René-Marie Lafontaine, R. Paul Scofield, Yingqiang Lou, Lankani Somarathna, Denis Lepage, Marshall Illif, Eike Lena Neuschulz, Mathias Templin, D. Matthias Dehling, Jacob C. Cooper, Olivier S. G. Pauwels, Kangkuso Analuddin, Jon Fjeldså, Nathalie Seddon, Paul R. Sweet, Fabrice A. J. DeClerck, Luciano N. Naka, Jeffrey D. Brawn, Alexandre Aleixo, Katrin Böhning-Gaese, Carsten Rahbek, Susanne A. Fritz, Gavin H. Thomas, and Matthias Schleuning. 2022. AVONET: morphological, ecological and geographical data for all birds. *Ecology Letters* 25, 3 (2022), 581–597. https://doi.org/10.1111/ele.13898 arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/ele.13898

[23] Mohamed Wardeh, Mark S. C. Blagrove, Kieran J. Sharkey, Matthew Baylis, and James L. N. Wood. 2021. Divide-and-conquer: machine-learning integrates mammalian and viral traits with network features to predict virus-mammal associations. *Nature Communications* 12 (2021), 3954. https://doi.org/10.1038/s41467-021-24085-w

[24] Mohamed Wardeh, Claire Risley, Megan K. McIntyre, Christoph Setzkorn, and Matthew Baylis. 2015. Database of host–pathogen and related species interactions, and their global distribution. *Scientific Data* 2 (2015), 150049. https://doi.org/10.1038/sdata.2015.49

[25] Hamish Wilman, Jonathan Belmaker, Jennifer Simpson, Carolina de la Rosa, Marcelo M. Rivadeneira, and Walter Jetz. 2014. EltonTraits 1.0: Species-level foraging attributes of the world's birds and mammals. *Ecology* 95 (2014), 2027–2027.

[26] Zonghan Wu, Shirui Pan, Fengwen Chen, Guodong Long, Chengqi Zhang, and Philip S. Yu. 2021. A Comprehensive Survey on Graph Neural Networks. *IEEE Transactions on Neural Networks and Learning Systems* 32, 1 (2021), 4–24. https://doi.org/10.1109/TNNLS.2020.2978386

[27] Jie Zhou, Ganqu Cui, Shengding Hu, Zhengyan Zhang, Cheng Yang, Zhiyuan Liu, Lifeng Wang, Changcheng Li, and Maosong Sun. 2020. Graph neural networks: A review of methods and applications. *AI Open* 1 (2020), 57–81. https://doi.org/10.1016/j.aiopen.2021.01.001

CSE 8803 EPI Project Proposal

Akshath Shvetang Anna
aanna7@gatech.edu
Georgia Institute of Technology
Atlanta, Georgia, USA

Grace Driskill
gdriskill3@gatech.edu
Georgia Institute of Technology
Atlanta, Georgia, USA

Julia Qian
jqian310@gatech.edu
Georgia Institute of Technology
Atlanta, Georgia, USA

Abstract

A virus's host range is an important indicator of its zoonotic potential. Predicting the missing hosts of viruses is thus an important step in supporting zoonotic disease surveillance, outbreak prevention and outbreak response. Here, we survey current methods and approaches in predicting missing links between animal species and pathogens. We then propose an ensemble learning model to predict associations between bird species and viruses, using an aggregated dataset with bird and virus features.

ACM Reference Format:
Akshath Shvetang Anna, Grace Driskill, and Julia Qian. 2018. CSE 8803 EPI Project Proposal. In *CSE 8803 EPI, Fall, 2024, Atlanta, GA.* ACM, New York, NY, USA, 4 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

A zoonotic disease is an infectious illness transmitted from animals to humans. Approximately 60% of all human illnesses originate from animals [8], including high-profile diseases such as Ebola and COVID-19 (both believed to originate from bats). The WHO estimates that approximately 1 billion cases of zoonoses occur annually, with millions of deaths [13]. Avian zoonoses, originating from birds, present a unique challenge regarding surveillance and predictions. Birds, like mammals, can act as hosts to viruses. Alternatively, they can spread the disease to vectors which then pass the infection to humans (such as the bird-mosquito-human pathway for West Nile virus). Moreover, birds can greatly increase the geographic range of a disease by spreading it anywhere along their migration routes [15], infecting both humans and domesticated animals.

Spillover is the event in which a disease isolated to one species begins infecting another. The range of species a virus can infect (host plasticity) is directly related to its potential to spillover to humans and its geographic reach [9]. Predicting a virus's host range can grant insight into which disease may become zoonotic, helping forecast potential outbreaks and supporting surveillance efforts.

Research efforts in this area by Wardeh et al. [17] have leveraged machine learning to predict previously unknown connections between mammal species and known viruses. Their results reveal over 20,000 potential unknown associations between viruses and mammals, pointing to how small a fraction of virus-mammal associations are known to researchers. Although their model provides extensive predictions for mammal hosts, with 1.67 million undiscovered viruses that exist in mammals and bird hosts [2], this remains an incomplete picture. This project aims to analyze host-pathogen and interspecies avian datasets to predict unknown associations between birds and viruses, granting better insight into a virus's range of hosts and by extension their potential to spill over to humans.

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM
https://doi.org/XXXXXXX.XXXXXXX

2 Survey of Related Work

Given the global health burden presented by zoonotic disease and the relationship between a virus's host range and its zoonotic potential [9], there has been an increasing interest in understanding the full scope of a virus's existing and potential hosts.

Farrell et al. [4] aimed to predict unknown associations between a global network of mammals and parasites by leveraging network attributes and phylogenetic relationships. They used a three-prong approach to predict links between hosts and parasites using observed interactions, using phylogenetic history, and combining both. They were able to effectively predict missing associations between mammals and parasites with sparse network data. One limitation of their approach is that, by using interaction data, their results may be biased due to targeted research. Associations may be over or under-reported due to factors like geographical accessibility (i.e. it may be harder to reach some populations of animals for testing), or an increased interest in certain mammals due to current outbreaks. They attempt to mitigate this bias by introducing phylogenetic and combined methodologies. We can adopt a similar approach by incorporating data sources less susceptible to reporting biases, such as phylogenetic or genomic data, to construct our network analysis. An additional limit is that their model did not include geographic features, meaning their model would predict interactions that, in the real world, are unlikely due to geographic and ecological distance. This can be addressed with the introduction of geographic data as a feature such that virus-mammal pairs spanning large distances are less likely to occur.

Similarly, Wardeh et al. [17] predict the missing associations between mammals and viruses through ensemble learning. They created three machine learning models to predict virus-mammal associations, mammal-virus associations, and missing edges in a global bipartite network, before applying ensemble voting. Although their paper is focused on mammal-virus associations, their goal to predict unknown animal-virus associations provides insight into how we can approach our assessment of avian-virus associations. However, like Farrell et al. [4], their results are biased by the research efforts of each virus. Their global network view provides an alternative method to deal with these research and sampling biases.

Olival et al. [11] investigate factors that impact a virus's host range and transmissibility to other species. They discover that viral richness, the number of different viruses found in a host species, is impacted by research effort, mammal sympatry (when two species

live in the same area) and host geographic range. In turn, factors that impact a mammal species's zoonotic viral richness include overall viral richness, phylogenetic distance from humans, host order and research effort. Although their research targets human-virus associations, their research highlights important geographical and phylogenetic factors that impact a virus's transmission from one species to another. We can apply the features they found that impact viral richness to our model to better estimate how many viruses different avian-species host. Similar to the previous two papers, their work is also potentially biased by the attention each virus receives in the research space and cannot account for associations that currently remain undiscovered.

Work by Elmasri et al. [3] used non-linear scaling of phylogenetic data to predict inter-species interactions. They applied a hierarchical Bayesian model for efficient link prediction in ecological networks. They suggest that their model can be applied to identify interactions that suffer from research biases (i.e. they are over-reported in research), offering an alternative solution to limits from Farrell et al. [4], Wardeh et al. [17], and Olival et al. [11]. Their approach is limited by the lack of diversity of data and model performance may improve through the inclusion of additional features such as geographical data, mammal sympatry, or species traits.

Poisot et al. [14] address the challenge of predicting missing host-virus interactions as a network problem. They represent the network as a bipartite structure, where virus nodes connect to host nodes where an association is known. They used a combined linear filtering and singular value decomposition model with hyper-parameters for determining the importance of different network structure characteristics. By choosing hyper-parameters that emphasize the network's connectance instead of in-degree or out-degree, they helped to reduce the bias caused by uneven research efforts, where certain parts of the network have more data due to concentrated study. Their techniques lead to predictions for possible new host-virus interactions that are robust to these data biases and suggest that future efforts for virus discovery be focused on the Amazon Basin and sub-Saharan Africa. While their method effectively handled the issue of incomplete and unbalanced data, a limitation was the exclusion of host and viral traits as predictors, which could enhance the model's performance.

Altizer et al. [1] investigates the relationship between animal migration and zoonotic diseases, suggesting that migration can both facilitate and mitigate disease spreading. A common concern is that migratory species may move diseases long distances geographically, and although this is a possibility, there are few documented examples of this phenomenon. More likely concerns are that the migratory species themselves are exposed to a wider range of diseases throughout their migration and that hotspots for diseases are created in places where multiple species of migratory animals stop to rest. An example of this is Delaware Bay, where stopping shorebirds create a hotspot of avian influenza. The stress of migration may lower the animals' immune system making them more susceptible to disease. However, Altizer et al. [1] also suggest that migration can actually decrease the spread of disease because staying in one habitat for a long period of time allows parasites to accumulate. Additionally, migration removes diseased members of a species from a community, and then decreases the chances of them spreading the illness, because they are not able to complete the migration.

This information is important when considering what factors to take into account when trying to predict potential zoonotic diseases. Due to the complicated role that migration has in zoonotic disease transmission, including migratory animals and information on their migrations could impact the model.

3 Problem Formulation

The problem we aim to address is that, while some associations between hosts and viruses are known, many associations between hosts and known viruses remain unknown. Using a dataset of bird traits, a dataset of virus traits and a dataset of bird-viruses associations, our goal is to extract key features from each to predict these missing associations between birds and viruses. The models we create will work from three independent perspectives (avian perspective, viral perspective, and network perspective) and combine using majority voting to minimize biases of each data source. In other words, we will predict which birds are potential hosts to known viruses, with an emphasis on predicting previously unreported relationships.

4 Data

For this project, we plan to use the datasets used in Wardeh et al. [17] supplemented with additional data on avian traits and avian viruses from various sources that are freely available online.

To get virus-bird associations, we plan to use ENHanCEd Infectious Diseases Database (EID2)[18], the database used in Wardeh et al. [17] for virus-mammal associations, along with the Bacterial and Viral Bioinformatics Resource Center (BV-BRC) [12].

BV-BRC has 339,347 results of virus where the host group is filtered by avian. However, there are multiple entries for each virus-host pair because there is an entry for each strain of the virus. After consolidating the entries, there will be a reduced number of host-virus associations. Results of searches from BV-BRC can be downloaded as CSV, text or Excel files.

EID2 has 13,676 results for filtering their organisms interactions page by filtering the carrier taxa to vertebrates and cargo taxa to viruses. From here, we will have to figure out which of the results belong to the class Aves because vertebrates encompass many taxonomy classes. One possible technique for this is using the National Center for Biotechnology Information (NCBI) dataset's API to search their taxonomy dataset by the carrier's name or taxid, so that we can determine the carrier's class. Then, we will have to get rid of overlapping host-virus interactions already found from the BV-BRC search. There is a note on this database's webpage to contact the EID2 team for requests to extract large volumes of data. At this point, we have not contacted the EID2 team yet, so we do not know the exact format of the data. However, it is likely to be a CSV or Excel file as the data is displayed in a table on the web interface.

The exact count of bird-virus associations we will extract from these sources is not known at this time, before the data processing, but will be less than the initial search result counts.

The traits for each bird will include phylogenetic, ecological and geospatial traits, mirroring the mammal traits used in Wardeh et al. [17]. We plan to use the EltonTraits 1.0 [19] dataset to get information on birds' diet and body mass, AVONET [16] to get

information on birds' geographical range, migration, trophic level and niche, and primary lifestyle, and BirdTree [7] to get birds' phylogeny and taxonomy.

Both EltonTraits and AVONET datasets can be downloaded online as CSV files. EltonTraits has data for 9,993 extant bird species; AVONET has data for 11,009 extant bird species. BirdTree data can be downloaded online as TRE files, and has data for 9,993 species. The exact amount of data entries we will use from these sources will depend on how many unique species are found from the bird-virus association datasets.

The viral traits and sources for the avian viruses will also mirror those used in Wardeh et al. [17]. The NCBI Taxonomy Database [5] will provide classifications of the virus: whether the virus is RNA-based, retro-transcribing, and whether it is negative-sense or positive-sense. ViralZone [6] will be used for gnomic characteristics: whether the viral genome is circular or linear, monopartite or segmented, if the virus is enveloped, GC-content, and average genome size. Both ViralZone and the ICTV [10] database will be used to get information on replication sites, viral release mechanisms, and methods of cell entry.

The NCBI Taxonomy Database can either be downloaded as .dmp files or accessed through their API. ViralZone and ICTV do not have API access or a way to download data, so for these sources we will mine the relevant information from the website and put it into a more usable format. Similar to bird traits, the exact amount of data from these sources will depend on how many unique viruses are from the bird-virus association datasets.

5 Approach

Our general approach will be modeled after the methodology of separating perspectives as demonstrated by Wardeh et al. [17]. In their study, they independently analyzed and trained models along three different perspectives:

(1) a viral perspective that explored which potential mammalian hosts each virus could infect, based on the host characteristics
(2) a mammalian perspective that explored which viruses could potentially infect each mammal based on the viral characteristics
(3) a network perspective that is organized as a bipartite graph with viral and mammalian sets

The last perspective captures global connectivity between viruses and mammals, and topological features are extracted from this network to use as features to predict associations between individual viruses and hosts.

The combination of these three perspectives allowed the researchers to counteract biases in the available data and to arrive at more accurate predictions. We plan to follow the same structure, substituting mammalian features for avian features and exploring an avian perspective instead of a mammalian one.

Additionally, the first two perspectives are built by training local suites of models for each pairwise combination of host and virus according to each perspective. Wardeh et al. [17] used eight different classification algorithms as part of the suite. Due to time constraints, we will limit our training to the top three performing algorithms based on how often they were chosen as the best of the eight. Based on the supplemental data, for the mammalian perspective, these were Support Vector Machines (SVM) with radial basis kernel and class weights (SVM-RW), SVM with linear kernel and class weights, and Random Forest. For the viral perspective, these were eXtreme gradient boosting, Naive Bayes, and Stochastic Gradient Boosting. It is possible that these will not be the best performing for the avian data, but they represent a useful starting point. The Synthetic Minority Over-sampling Technique (SMOTE) may also be used to address class imbalance between the average number of hosts each virus affected and the average number of viruses each host was affected by. This was necessary with the mammalian data, and it remains to be seen whether it will be for the avian data we are organizing.

The models were trained with 10-fold cross-validation, and this will be carried over. 50 replicates of each were also made to allow for the generation of a 90% confidence interval. Due to time constraints, this will also be downsized to just 10 replicates.

For the network perspective, since there were relatively few known host-virus associations amongst the large number of possible ones, Wardeh et al. [17] balanced the known associations and possible yet unknown associations by randomly choosing 1000 unknown associations and under-sampling the known associations. We will likely face a similar imbalance and would follow the same procedure. 10-fold cross validation was then used to train the eight algorithms from before with 100 replicate models for estimating confidence intervals which were aggregated into ensemble models, with the best overall performer across all the output metrics being chosen. Their best performer was SVM-RW. Therefore, with 20 replicates we will start with this algorithm and explore the potential of the others as time allows.

6 Evaluation

To evaluate our models, we will use a combination of the Area Under the Receiver Operating Characteristic Curve (AUC), true skills statistics (TSS), and F-1 score. These were the metrics used by Wardeh et al. [17] and we will use the same to allow for easy comparison to that study.

7 Expectation

By the end of the semester, we expect to create three datasets: one of avian traits, one of viral traits and one of avian-virus associations. We will use these datasets to train models from three perspectives: the avian perspective, the viral perspective and the network perspective. Finally, we will combine these three perspectives to predict missing host-virus associations.

8 Work Distribution

All three members will split the work evenly throughout this project. We will have weekly meetings to update each other on our progress and communicate concerns or questions. Furthermore, we will also seek clarification from the professor or TA for any questions we come across during our development process.

Initially, data aggregation and cleaning will be done primarily by two team members (Grace Driskill, Julia Qian), while the remaining member (Akshath Anna) begins to investigate the different machine learning models we intend to train. Models will be implemented

in Python. We believe that data aggregation and cleaning will be a highly involved process since we need to aggregate and normalize data across several datasets. Thus, working on the data and machine learning models in parallel will allow us to begin running the models with minimal delay after the dataset is processed. To ensure consistency between the extraction of features from the dataset and our machine learning models, we will maintain open communication and documentation between both sides. After the data has been cleaned, all three of us will focus on completing the machine learning models, fine-tuning parameters, generating predicted associations, and validating our results. Additionally, we will contribute equally to the midterm report, final report and final presentation.

9 Expected Timeline

There are approximately four weeks before the Project Milestone and then four more weeks before the final presentation. In the first two weeks, we will work in parallel to clean and aggregate the data and begin developing our machine learning model in Python. Two team members will focus on generating the avian dataset by aggregating entries across several data sources. The other team member will investigate and begin implementing our models in Python. After the dataset is constructed, throughout the next three weeks, all teammates will continue working on developing our machine learning model. We will then begin training and testing the model on our avian dataset. Here, we will continue to engineer features to improve our model's performance. In the following two weeks, we will evaluate the performance of our model and begin working on the final report and presentation. The remaining week will act as buffer time to deal with any issues we run into as we build, train, and evaluate the model.

References

[1] Sonia Altizer, Rebecca Bartel, and Barbara A. Han. 2011. Animal Migration and Infectious Disease Risk. *Science* 331, 6015 (2011), 296–302. https://doi.org/10.1126/science.1194694 arXiv:https://www.science.org/doi/pdf/10.1126/science.1194694

[2] Dennis Carroll, Peter Daszak, Nathan D Wolfe, George F Gao, Carlos M Morel, Subhash Morzaria, Ariel Pablos-Méndez, Oyewale Tomori, and Jonna AK Mazet. 2018. The global virome project. *Science* 359, 6378 (2018), 872–874.

[3] Mohamad Elmasri, Maxwell J Farrell, T Jonathan Davies, and David A Stephens. 2020. A hierarchical Bayesian model for predicting ecological interactions using scaled evolutionary relationships. *Annals of Applied Statistics* (2020).

[4] Maxwell J Farrell, Mohamad Elmasri, David A Stephens, and T Jonathan Davies. 2022. Predicting missing links in global host–parasite networks. *Journal of Animal Ecology* 91, 4 (2022), 715–726.

[5] Scott Federhen. 2012. The NCBI Taxonomy database. *Nucleic Acids Research* 40 (2012), D136–D143. https://doi.org/10.1093/nar/gkr1178

[6] Christophe Hulo, Edouard de Castro, Pierre Masson, Lydie Bougueleret, Amos Bairoch, Ioannis Xenarios, and Philippe Le Mercier. 2011. ViralZone: a knowledge resource to understand virus diversity. *Nucleic Acids Research* 39 (2011), D576–D582. https://doi.org/10.1093/nar/gkq901

[7] Walter Jetz, Gavin H. Thomas, Jeffrey B. Joy, Klaas Hartmann, and Arne O. Mooers. 2012. The global diversity of birds in space and time. *Nature* 491 (2012), 444–448.

[8] Kate E Jones, Nikkita G Patel, Marc A Levy, Adam Storeygard, Deborah Balk, John L Gittleman, and Peter Daszak. 2008. Global trends in emerging infectious diseases. *Nature* 451, 7181 (2008), 990–993.

[9] Christine Kreuder Johnson, Peta L Hitchens, Tierra Smiley Evans, Tracey Goldstein, Kate Thomas, Andrew Clements, Damien O Joly, Nathan D Wolfe, Peter Daszak, William B Karesh, et al. 2015. Spillover and pandemic properties of zoonotic viruses with high host plasticity. *Scientific reports* 5, 1 (2015), 14830.

[10] Elliot J. Lefkowitz, Donna M. Dempsey, Robert C. Hendrickson, Richard J. Orton, Stuart G. Siddell, and David B. Smith. 2018. Virus taxonomy: The database of the International Committee on Taxonomy of Viruses (ICTV). *Nucleic Acids Research* 46 (2018), D708–D717. https://doi.org/10.1093/nar/gkx932

[11] Kevin J Olival, Parviez R Hosseini, Carlos Zambrana-Torrelio, Noam Ross, Tiffany L Bogich, and Peter Daszak. 2017. Host and viral traits predict zoonotic spillover from mammals. *Nature* 546, 7660 (2017), 646–650.

[12] Robert D. Olson, Rima Assaf, Thomas Brettin, Neil Conrad, Cody Cucinell, James J. Davis, Donna M. Dempsey, Alan Dickerman, Elizabeth M. Dietrich, Robert W. Kenyon, Melis Kuscuoglu, Elliot J. Lefkowitz, James Lu, Danielle Machi, Catherine Macken, Chunhong Mao, Anna Niewiadomska, Maria Nguyen, Gary J. Olsen, Ross C. Overbeek, Bryan Parrello, Vincent Parrello, Joshua S. Porter, Gordon D. Pusch, Maulik Shukla, Indu Singh, Leon Stewart, Grace Tan, Cathy Thomas, Mary VanOeffelen, Veronika Vonstein, Zachary S. Wallace, Alexander S. Warren, Alice R. Wattam, Fangfang Xia, Haechan Yoo, Yan Zhang, Christian M. Zmasek, Richard H. Scheuermann, and Rick L. Stevens. 2023. Introducing the Bacterial and Viral Bioinformatics Resource Center (BV-BRC): a resource combining PATRIC, IRD and ViPR. *Nucleic Acids Research* 51, D1 (2023), D678–D689. https://doi.org/10.1093/nar/gkac1003

[13] World Health Organization. [n. d.]. WHO EMRO: | Zoonotic disease: emerging public health threats in the Region. https://www.emro.who.int/about-who/rc61/zoonotic-diseases.html/

[14] Timothée Poisot, M. Annie Ouellet, Nicolas Mollentze, Matthew J. Farrell, Daniel J. Becker, Leonard Brierley, Gregory F. Albery, Rory J. Gibb, Stephanie N. Seifert, and Colin J. Carlson. 2023. Network embedding unveils the hidden interactions in the mammalian virome. *Patterns (N Y)* 4, 6 (2023), 100738. https://doi.org/10.1016/j.patter.2023.100738

[15] Kurt D Reed, Jennifer K Meece, James S Henkel, and Sanjay K Shukla. 2003. Birds, migration and emerging zoonoses: West Nile virus, Lyme disease, influenza A and enteropathogens. *Clinical medicine & research* 1, 1 (2003), 5–12.

[16] Joseph A. Tobias, Catherine Sheard, Alex L. Pigot, Adam J. M. Devenish, Jingyi Yang, Ferran Sayol, Montague H. C. Neate-Clegg, Nico Alioravainen, Thomas L. Weeks, Robert A. Barber, Patrick A. Walkden, Hannah E. A. MacGregor, Samuel E. I. Jones, Claire Vincent, Anna G. Phillips, Nicola M. Marples, Flavia A. Montaño-Centellas, Victor Leandro-Silva, Santiago Claramunt, Bianca Darski, Benjamin G. Freeman, Tom P. Bregman, Christopher R. Cooney, Emma C. Hughes, Elliot J. R. Capp, Zoë K. Varley, Nicholas R. Friedman, Heiko Korntheuer, Andrea Corrales-Vargas, Christopher H. Trisos, Brian C. Weeks, Dagmar M. Hanz, Till Töpfer, Gustavo A. Bravo, Vladimír Remeš, Larissa Nowak, Lincoln S. Carneiro, Amilkar J. Moncada R., Beata Matysioková, Daniel T. Baldassarre, Alejandra Martínez-Salinas, Jared D. Wolfe, Philip M. Chapman, Benjamin G. Daly, Marjorie C. Sorensen, Alexander Neu, Michael A. Ford, Rebekah J. Mayhew, Luis Fabio Silveira, David J. Kelly, Nathaniel N. D. Annorbah, Henry S. Pollock, Ada M. Grabowska-Zhang, Jay P. McEntee, Juan Carlos T. Gonzalez, Camila G. Meneses, Marcia C. Muñoz, Luke L. Powell, Gabriel A. Jamie, Thomas J. Matthews, Oscar Johnson, Guilherme R. R. Brito, Kristof Zyskowski, Ross Crates, Michael G. Harvey, Maura Jurado Zevallos, Peter A. Hosner, Tom Bradfer-Lawrence, James M. Maley, F. Gary Stiles, Hevana S. Lima, Kaiya L. Provost, Moses Chibesa, Mmatjie Mashao, Jeffrey T. Howard, Edson Mlamba, Marcus A. H. Chua, Bicheng Li, M. Isabel Gómez, Natalia C. García, Martin Päckert, Jérôme Fuchs, Jarome R. Ali, Elizabeth P. Derryberry, Monica L. Carlson, Rolly C. Urriza, Kristin E. Brzeski, Dewi M. Prawiradilaga, Matt J. Rayner, Eliot T. Miller, Rauri C. K. Bowie, René-Marie Lafontaine, R. Paul Scofield, Yingqiang Lou, Lankani Somarathna, Denis Lepage, Marshall Illif, Eike Lena Neuschulz, Mathias Templin, D. Matthias Dehling, Jacob C. Cooper, Olivier S. G. Pauwels, Kangkuso Analuddin, Jon Fjeldså, Nathalie Seddon, Paul R. Sweet, Fabrice A. J. DeClerck, Luciano N. Naka, Jeffrey D. Brawn, Alexandre Aleixo, Katrin Böhning-Gaese, Carsten Rahbek, Susanne A. Fritz, Gavin H. Thomas, and Matthias Schleuning. 2022. AVONET: morphological, ecological and geographical data for all birds. *Ecology Letters* 25, 3 (2022), 581–597. https://doi.org/10.1111/ele.13898 arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/ele.13898

[17] Mohamed Wardeh, Mark S. C. Blagrove, Kieran J. Sharkey, Matthew Baylis, and James L. N. Wood. 2021. Divide-and-conquer: machine-learning integrates mammalian and viral traits with network features to predict virus-mammal associations. *Nature Communications* 12 (2021), 3954. https://doi.org/10.1038/s41467-021-24085-w

[18] Mohamed Wardeh, Claire Risley, Megan K. McIntyre, Christoph Setzkorn, and Matthew Baylis. 2015. Database of host–pathogen and related species interactions, and their global distribution. *Scientific Data* 2 (2015), 150049. https://doi.org/10.1038/sdata.2015.49

[19] Hamish Wilman, Jonathan Belmaker, Jennifer Simpson, Carolina de la Rosa, Marcelo M. Rivadeneira, and Walter Jetz. 2014. EltonTraits 1.0: Species-level foraging attributes of the world's birds and mammals. *Ecology* 95 (2014), 2027–2027.