

Predicting unknown viral hosts with Dynamic Positive-Unlabeled learning

Supplementary Information

Gabriele Pignalberi¹, Andre Tonelli², Stefano Giagu¹, and Moreno Di Marco²

¹*Department of Physics, Sapienza University of Rome, Rome, Italy*

²*Department of Biology and Biotechnologies 'Charles Darwin', Sapienza University of Rome, Rome, Italy*

Contents

Supplementary Methods	2
1 Input Data	2
1.1 Mammal species data	2
1.2 Virus data	3
1.3 Synthetic data	3
2 Modeling Approach	4
2.1 Observational biases	4
2.2 Modeling relationships between data	5
2.3 Dynamic Positive-Unlabeled framework	6
2.4 Bayesian DPU model	9
2.5 Classifier model	15
3 Model Training	18
3.1 Dynamic Positive-Unlabeled learning	18
3.2 Graph-dataset creation	20
3.3 Our training approach	21
4 Uncertainty calculation, aggregated predictions and limitations	23
4.1 Uncertainty calculation	23
4.2 Aggregated predictions	24
4.3 Limitations of uncertainty estimates	25
Supplementary Results	26
1 Predictions for each mammalian orders	26
2 Predictions for each viral families	39
2.1 Predictions across mammalian orders	39
2.2 Predicted geographic distribution of the hosts	73
3 Propensity scores and uncertainties	106
4 Predictions by mammal degrees	107
5 Table of predictions	107
Supplementary Notes	109
1 Random continuous variables in $[0, 1]$	109
2 Positive-Unlabeled measures	109
2.1 Metrics for PU learning	110
2.2 Extending PUF_1 to SAR scenario	110
2.3 Naive evaluation metrics in PU scenarios	111
3 Pseudocode of the algorithm for partitioning associations	112
Supplementary Information References	113

Supplementary Methods

1 Input Data

This section offers supplementary details about the input data used in our research, in addition to those presented in the article’s methods section.

1.1 Mammal species data

We consider 5330 terrestrial mammal species $m \in M$. These are all the mammalian species for which we have detailed information about their life-history traits, phylogenetic relationships with other species, and geographic range. We followed mammalian taxonomy as reported in the IUCN Red List v. 2020-2 [1], as this is the same taxonomy adopted in the VIRION database.

Mammalian features

Each mammal species $m \in M$ is associated to vector \mathbf{x}_m of 18 features. Here, we offer a brief description of the types of mammalian features used (full details available in Supplementary Table SM1):

- **Morphological traits.** Morphological traits are observable and measurable physical features of an organism’s phenotype. Also, morphological traits are known to exhibit correlations with host-pathogen associations [2].
- **Life-history traits.** Life-history traits are the biological characteristics and strategies of an organism, selected over evolutionary time, that influence its growth, reproduction, social behaviors, lifespan and survival. These traits can modulate the host’s susceptibility to various classes of pathogens and impact the transmission dynamics of pathogens within host populations by affecting factors such as host density, contact rates, and immune responses [3].
- **Ecological traits.** Ecological traits, such as diet, habitat preferences, and resource use, characterize how a species interacts with its environment and other organisms.
- **Phylogenetic eigenvectors.** Following COMBINE [4], in order to include phylogenetic information, we randomly selected 1 phylogeny from PHYLACINE v. 1.2 database [5] (ID: 1), and extracted the first 10 eigenvectors. The entries of these phylogenetic eigenvectors defines the phylogenetic spectral embeddings. These allows us to map mammalian species onto a low-dimensional space where ‘phylogenetically close’ species are positioned close also in terms of their distance in the embedding space.

Phylogeny can help explaining the inheritance and lineage of traits or specific genes among species. This makes phylogenetic relatedness a strong predictor of viral sharing across species, as it reflects the species patterns of genetic diversity and the potential co-evolution of pathogens and hosts.

We note that spectral eigenvectors are always defined up to a sign. To address this multiplicity, machine learning models such as Graph Transformers [6] randomly flip the signs of spectral embeddings during the model’s training. In contrast, we fixed the sign of each eigenvector to define a consistent coordinate system across species. In fact, while flipping the sign of an eigenvector preserves pairwise distances—and thus retains phylogenetic relatedness—it alters species’ absolute positions in the embedding space. Instead, by fixing eigenvector signs, we establish a canonical coordinate system in which each species’ coordinates consistently reflect its absolute position within the phylogenetic tree. This allows us to preserve not only phylogenetic relatedness but also taxonomic information, which is related with species’ absolute positions in the tree.

All mammalian features were normalized to have a feature-wise mean of zero and a standard deviation of one before being processed by our models.

Mammal pairwise phylogentic distances

To model dynamics of coevolution among mammalian species, we use patristic distances, $t(m, m')$, which represent the elapsed time since the most recent common ancestor of species m and m' . However, inferring phylogenies comes with challenges, including discrepancies in patristic distances due to various factors [9, 10]. In this study, we use the

Supplementary Table SM1 | Mammalian features used and their description. Summary of morphological, life-history, and ecological traits used to characterize mammalian susceptibility to pathogens.

Feature type	Feature name	Feature description
Morphological traits	Log adult body mass	Proxies metabolism, adaptation, and immunity phenotypes [7], affecting species' susceptibility to pathogens [2].
Life-history traits	Maximum longevity	Long-lived species can maintain infections within the populations over extended periods. Also, the presence of different age groups may lead to varying level of immunity [2].
	Gestation length	Shorter gestation periods mean higher reproductive output and population densities.
	Interbirth interval	Influences pathogen spread rate; shorter intervals lead to rapid population growth.
	Log weaning age	Reflects trade-off between reproductive effort and survival, affecting resources available for immunity [8].
	Mean litter size	Indicates reproductive output; larger litters suggest faster life histories.
	Litter size for years	Annual average of reproductive output across multiple breeding events.
Ecological traits	Trophic level	Represents position in the food chain, from primary consumers (level 1) to apex predators (level 3).

PHYLACINE v. 1.2 database [5], which provides patristic distances from a posterior distribution of 1,000 phylogenetic trees. While these trees show minor variations for most species, they exhibit greater discrepancies for species with limited or no genetic data. However, as these variations are unlikely to significantly impact our predictions, we base our analysis using a single, randomly selected, phylogeny (ID: 1), from which we extract both patristic distances and phylogenetic eigenvectors (see above).

Mammalian biogeography

In addition to phylogeny, another crucial aspect for describing mammalian species is biogeography. Quantitative studies have demonstrated that biogeography is highly correlated with viral sharing within animal populations, making it, alongside phylogeny, one of the most significant predictors of viral sharing among mammalian species [11–15]. This correlation arises because biogeography reflects the physical interactions between sympatric species, which plays a fundamental role in the transmission of viruses. Additionally, species within the same biogeographic regions are usually subjected similar environmental conditions, such as climate and vector species, highlighting the importance of incorporating biogeography into our analysis.

Geographic ranges of mammalian species and their spatial overlaps $o(m, m')$ were sourced from the IUCN Red List v. 2020-2 database [1].

1.2 Virus data

We denote the set of all viral families considered in our work as V , with individual viral families represented by elements $v \in V$. V contains a total of 33 distinct viral families, out of the 40 known viral families reported to affect mammals in the VIRION database.

Viral features

Each viral family $v \in V$ is associated to vector \mathbf{x}_v of 7 viral features. All viral features were sourced from Wardeh et al. 2021 [16] and are detailed in Supplementary Table SM2.

1.3 Synthetic data

We simulated associations between mammalian species and two synthetic viral families, for which we controlled both the ground truth host ranges and the labeling mechanisms.

For both families, viral traits were randomly assigned by sampling observed combinations of features from the real dataset.

To simulate observational biases, we generate labeled instances from synthetic positive associations using synthetically constructed propensity scores. For a given synthetic viral family w , the synthetic propensity score $e(m, w)$ for mammal species m were defined as:

$$e(m, w) = \frac{k_{m,w} \tilde{e}(m)}{1 + (k_{m,v} - 1) \tilde{e}(m)} \quad (\text{SM1})$$

Supplementary Table SM2 | Viral features used and their description. Summary of genetic traits used to characterize host range of the viral families.

Viral feature	Feature description
RNA/DNA	This is a binary variable where 1 represents RNA viruses and 0 DNA viruses. RNA viruses mutate and adapt faster than DNA viruses [17] and typically deactivate quickly in the environment. Compared to DNA viruses, RNA viruses are less dependent on host phylogeny [15, 18, 19].
Retro-transcribing	A binary variable indicating whether the virus is retro-transcribing. Retroviruses are highly conserved [20] and must integrate into the host genome by accessing the cell nucleus [21], limiting their target host range.
Strand structure	A multiple binary indicator that represents whether the virus is <i>single-stranded and positive-oriented</i> , <i>single-stranded and negative-oriented</i> , or <i>double-stranded</i> . These factors affect the viral replication cycle and interactions with host cellular machinery.
Circular/linear	A binary variable where 1 represents circular viruses and 0 represents linear ones. Circular genomes reduce the virus’ dependency on host enzymes for replication and translation [22], potentially affecting host range.
Segmented/monopartite	A binary variable with 1 representing segmented viruses and 0 representing monopartite viruses. Segmented viruses can recombine segments from different strains during co-infection, potentially resulting in new variants with altered properties, including host-switching capabilities [23, 24].
Envelope	A binary variable where 1 represents enveloped viruses and 0 non-enveloped ones. Viral envelopes help evade host immune systems but are sensitive to environmental factors, requiring direct transmission between hosts. Envelope composition can change to reflect new host cell membranes, increasing adaptability [25].

where

$$\tilde{e}(m) = \frac{1}{|V|} \sum_{v \in V} \hat{e}_{m,v}, \quad (\text{SM2})$$

is the average estimated propensity score—as defined in Supplementary Equation SM58—for species m across all real viral families $v \in V$. The factor $k_{m,w}$ is a user-defined scaling parameter that modulates observational bias and varies depending on the specie m and synthetic viral family w .

For the first synthetic viral family (Syn1), we set $k_{m,w} = 3$ for species in the United States, and $k_{m,w} = 1$ elsewhere, to simulate a marked geographic sampling bias. This procedure resulted in 32 labeled and 50 unlabeled positives. For the second synthetic viral family (Syn2), we set ($k_{m,w} = 3$) for ungulates (orders Cetartiodactyla and Perissodactyla), and $k_{m,w} = 1$ for all others, to simulate a marked taxonomic sampling bias. This procedure resulted in 201 labeled and 635 unlabeled positives.

These synthetic propensity scores are used solely for generating synthetic labels and are not provided to the model during either the training or testing steps. Instead, the estimated propensity scores used for training the classifier and correcting predictions via Supplementary Equation SM12 are obtained through the Bayesian modeling approach described in Section 2.4.

2 Modeling Approach

This section offers supplementary information on the strategy used to model observational data and predict missing associations.

- Subsection 2.1 addresses the major observational biases characterizing documented mammal–virus associations.
- Subsection 2.2 describes the relationships known among mammalian species, and those observed between mammalian species and viral families.
- Subsection 2.3 presents a novel and generalizable approach for formalizing the problem, which is the Dynamic Positive-Unlabeled (DPU) framework.
- Subsection 2.4 details the initialization of this framework via a hierarchical Bayesian DPU model.
- Subsection 2.5 introduces a classifier based on Graph Neural Networks (GNNs), designed to integrate all relevant factors for producing accurate predictions of mammal–virus association existence.

2.1 Observational biases

Research efforts have unevenly mapped the landscape of mammal–virus associations, resulting in concentrated knowledge around a limited subset of hosts and viruses, while leaving many other interactions under-described. As outlined

by Wardeh et al. (2021) [16], several key factors, often interrelated, can contribute to this imbalance. Here, we will name the most important ones.

1. **Human-centric biases.** Human pathogens, along with those affecting domesticated animals and livestock, attract the highest research attention and drive pathogen discovery efforts. As a result, current knowledge we have on the subject is hugely skewed towards *Homo sapiens* and domesticated animals, with far less attention given to wild species. Similarly, pathogens that are known to pose direct threats to human health or affect economically important livestock receive significant more attention.
2. **Geographic biases.** Virological sampling and research efforts are uneven across geographic space due to public health priorities, resource availability, and logistical constraints [26, 27]. Infectious diseases in rare or remote species are less frequently documented, given the logistical challenges of conducting surveillance and diagnostic campaigns in their habitats, as well as their distance from human settlements. Additionally, identifying host species for a virus tends to shift focus toward geographically co-occurring species. As a result, research efforts are often clustered around regions with better infrastructure, higher accessibility, and existing surveillance programs, leaving vast areas—particularly in biodiversity-rich regions—underrepresented in observed host–pathogen interactions.
3. **Biological biases.** Host–pathogen interactions that result in more noticeable illness or host mortality are more likely to be detected. This may potentially bias our understanding of a pathogen’s host-range by overlooking asymptomatic species that may play a key role in its maintenance in the ecosystem [28].
4. **Self-reinforcing biases.** Hosts that are known to be susceptible to certain types of pathogens are also more likely to be studied when a new viral agent of that type is found. Examples include coronaviruses for bats, or hantaviruses for rodents. This focus can distort our understanding of viral diversity and host-pathogen interactions by emphasizing researches on well-studied hosts and their associated viruses, while leaving a large proportion of potentially important hosts of these viruses underrepresented.

2.2 Modeling relationships between data

To estimate the true prevalence of viral families among mammalian species, our modeling approach should incorporate all the relevant factors necessary for a comprehensive characterization of both the ground truth and labeling mechanisms. This includes capturing both the complex relationships incurring among mammalian species and the observed network of mammal–virus associations. Such modeling is essential for uncovering eco-evolutionary patterns of mammal-virus associations and to produce accurate, reliable and biologically informed predictions. These relationships can be formalized using graph theory.

Graph of observed mammal-virus associations

Current knowledge of global virus sharing among mammalian hosts can be represented using the (bipartite and unweighted) graph of known mammal-virus associations $G_{MV} = (M \cup V, E_{MV})$. In Supplementary Figure SM1, we report the degree distribution for mammal species and viral families in G_{MV} (constructed using only labels from the VIRION database). This highlights the huge disproportion of research efforts across both mammalian species and viral families.

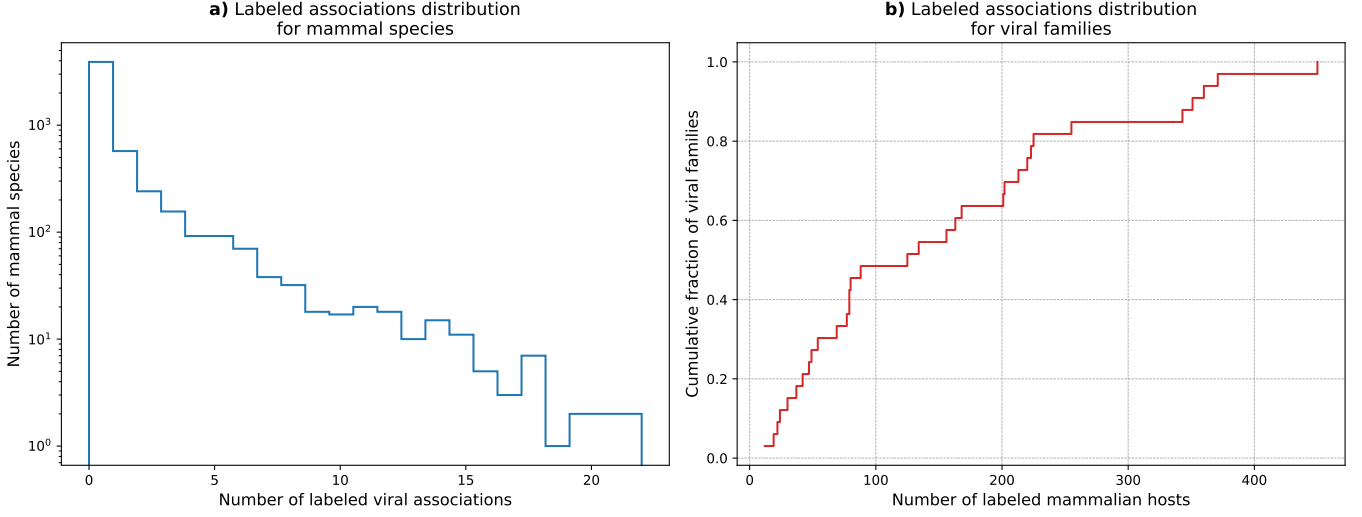
Phylogenetic and biogeographical mammalian graphs

Modeling phylogenetic and biogeographical relationships among mammals is crucial for uncovering latent patterns which could provide deeper insight into virus transmission pathways and compensate for observational data gaps by leveraging known interspecific relationships.

1. **Phylogenetic graph.** To model the dynamics of coevolution and phylogenetic similarities among mammalian species, we use patristic distances $t(m, m')$. We define a weighted adjacency operator $\mathbf{P}_\tau : M \times M \rightarrow [0, 1]$, which we refer to as the *mammalian phylogenetic adjacency matrix* subjected to *phylogenetic temperature* $\tau > 0$. Its matrix elements are given by:

$$\mathbf{P}_\tau[m, m'] = \frac{e^{-t(m, m')/\tau}}{\sum_{m'' \in M} e^{-t(m, m'')/\tau}}, \quad \forall m, m' \in M. \quad (\text{SM3})$$

The adjacency operator \mathbf{P}_τ , as defined, is normalized in a manner consistent with random-walk normalization and exhibits entries that decay exponentially as the original phylogenetic similarities between mammal species decrease. Although, we have $\mathbf{P}_\tau[m, m'] > 0 \forall m, m' \in M$, the exponential decay ensures that \mathbf{P}_τ is generally really sparse, in a sense it possesses many entries close to zero. This sparsity thereby emphasizes the most relevant phylogenetic connections while it dampens the less significant ones. Theoretically, multiple phylogenetic graphs can be generated at different phylogenetic temperatures to capture both short-range and long-range phylogenetic interactions between species.



Supplementary Figure SM1 | Distribution in labeled associations for mammalian species and viral families. **a)** The panel shows the histogram of the log-number of labeled associations (based on the VIRION database) with viral families for mammalian species. **b)** The panel shows the cumulative distribution of labeled associations (based on the VIRION database) with mammal species for viral families.

- Biogeographic graph (1-hop).** A natural approach to leverage mammalian biogeography is to represent the geographical adjacencies among species using a graph that encodes their spatial overlaps. Also, it is possible to incorporate edge weights to quantify *biogeographical influence* between the species. At this end, we define a weighted adjacency operator $\mathbf{G} : M \times M \rightarrow [0, 1]$, referred to as the (*mammalian*) *biogeographic adjacency matrix*. Its matrix elements are given by:

$$\mathbf{G}[m, m'] = \frac{o_{m, m'}}{\sum_{m''} o_{m, m''}}, \quad \forall m, m' \in M. \quad (\text{SM4})$$

where $o_{m, m'}$ is the spatial overlap between m and m' . The operator is clearly asymmetric, which is a desirable property. For instance, while *Mus musculus* (the house mouse) is found throughout the geographic range of *Lynx pardinus* (the Iberian lynx), the reverse is far from true. This makes the biogeographic influence of the Iberian lynx on the house mouse negligible in comparison with the influence of the house mouse on the Iberian lynx. Van Dam et al. (2021) [29] spectral clustering to biogeographic graphs of mammalian carnivores, revealing clusters that closely align with traditional biogeographic realms [30], highlighting the value of these graphs for analyzing macro-biogeographical patterns of the species.

- Biogeographic graph (2-hop).** Finally, we can also define the 2-hop biogeographic adjacency operator as:

$$\mathbf{G}^2[m, m'] = \sum_{m'' \in M} \mathbf{G}[m, m''] \mathbf{G}[m'', m'], \quad \forall m, m' \in M. \quad (\text{SM5})$$

This operator allows for biogeographic connections between mammal species that do not directly share the same geographic range but still are indirectly connected through a third species whose range overlaps with both. This approach extends biogeographic relationships beyond co-occurrence and direct geographic overlap. However, considering 2-hop geographic connections is particularly important. In fact, many species that belong to the same interconnected biogeographic cluster—and are therefore likely to share similar viral diversity and environment—may not have direct geographic overlap, but can still be well-connected through multi-hop biogeographic paths. This consideration is especially relevant for species in archipelagos, where physical separation by water can prevent direct contact between species, but proximity and ecological connectivity within the archipelago still facilitate viral sharing through intermediary species. A similar logic applies also to nearshore islands, where species on the island are often connected through 2-hop biogeographic links to species in the nearby coastal ecosystem.

2.3 Dynamic Positive-Unlabeled framework

In this Subsection, we analyze the task of predicting missing mammal-virus associations. This can be framed as a link-prediction problem on a finite (bipartite) network of associations, where only a biased subset of all the truly existent associations is reported.

Specific techniques are required for treating this type of problem. Here, we propose an innovative and generalizable framework, referred to as Dynamic Positive-Unlabeled (DPU), which extends the approach of traditional Positive-Unlabeled (PU) learning to systems with dynamic labels. Specifically, within a DPU approach, predictions are intended as inherently dynamic—that is, they adapt as more previously unlabeled instances become labeled.

DPU formalism

The characterization of a DPU problem follows multiple steps.

First, we assume that there exists an underlying *time-independent* structure $\Sigma = (\mathcal{V}, \Omega, Y, \mathbf{X}_g, \mathbf{X}_e)$ that defines the space of possibilities and the ground truth regarding our specific problem. This is defined by the following elements:

- \mathcal{V} represents the finite set containing all the *entities* considered for our problem. In our case, $\mathcal{V} = M \cup V$. The set \mathcal{V} is assumed as time-independent, that is, if an entity is an element of \mathcal{V} at a given time, it is also an element of it at all other times.
- Ω represents the finite set of all *possible associations* between entities. In our case, this set contains all the possible mammal-virus associations, i.e., $\Omega = M \times V$. However, more generally, we have $\Omega \subseteq \mathcal{V} \times \mathcal{V}$. Throughout the remainder of this work, we also refer to elements $x \in \Omega$ as instances.
- $Y : \Omega \rightarrow \{0, 1\}$ is a function that maps each association $x \in \Omega$ to its class $y \in \{0, 1\}$. An association with positive class (i.e., $y = 1$) is considered *true*, while a association with a negative class (i.e., $y = 0$) is considered *false*. Associated to Y , the set $\mathcal{P} = \{x \in \Omega : Y(x) = 1\} \subseteq \Omega$ represents the set of all true associations, while $\mathcal{N}(\mathbb{K}) = \{x \in \Omega : Y(x) = 0\} = \Omega \setminus \mathcal{P}$ is defined as the set of all false associations. In our specific context, \mathcal{P} corresponds to the set of all the truly existent mammal-virus associations in nature. The fundamental assumption of DPU learning is that \mathcal{P} is never directly and completely observed. For this reason, we can think of $y \in \{0, 1\}$ as a random variable, with $\Pr(y|x)$ representing the probability that association $x \in \Omega$ has class $Y(x) = y$. This reflects the epistemic uncertainty about the true class of x due to incomplete knowledge of Y .
- $\mathbf{X}_g : \mathcal{V} \rightarrow \mathbb{R}^d$ is a set of features that contains any type of relevant information for the characterization of entities in \mathcal{V} . We consider this information to be important for discriminating true associations from false ones and thus crucial for determining the ground truth mechanism. In our case, it represents the feature matrices \mathbf{X}_m and \mathbf{X}_v associated with mammalian species and viral families, respectively.
- $\mathbf{X}_e : \mathcal{V} \rightarrow \mathbb{R}^{d'}$ is a set of features that contains any type of relevant information *exclusively* associated with the entities' propensity to be studied. For our problem, we do not consider these features, however, for completeness we include them in the general formalism.

We also assume the existence of a *time-dependent* structure $\mathbb{K} = (S, \mathbf{X}_k)$, which represents a way to concisely characterize the configuration of our database at a given time, that is our current knowledge base on the subject. We refer to it as the *configuration* of our database at a given time. Also, for readability, we will not report its explicit dependence on the time. Generally, two different \mathbb{K} and \mathbb{K}' should be intended as representing two different configurations of the same database at two different times. Its elements are the following:

- $S : \Omega \rightarrow \{0, 1\}$ is a function that maps each association $x \in \Omega$ to its label $s \in \{0, 1\}$, as reported in the database at a given time. An association reported in the database (i.e., $s = 1$) is *labeled*, while an association which is not reported (i.e., $s = 0$) is *unlabeled*. This function provides a concise way of representing the labels of all possible associations at a given time. The assumption of DPU learning is that all labeled associations are true, i.e., $\mathcal{S}(\mathbb{K}) = \{x \in \Omega : S(x) = 1\} \subseteq \mathcal{P}$. As a result,

$$\Pr(y|x \in \mathcal{S}(\mathbb{K})) = \Pr(y|x, s = 1) = 1. \quad (\text{SM6})$$

Conversely, unlabeled associations can either be true or false. Finally, we define $\mathcal{U}(\mathbb{K}) = \{x \in \Omega : S(x) = 0\} = \Omega \setminus \mathcal{S}(\mathbb{K})$ as the set containing all the unlabeled instances in \mathbb{K} .

- $\mathbf{X}_k : \mathcal{S}(\mathbb{K}) \rightarrow \mathbb{R}^{d''}$ is a set of features that contains any other type of relevant information that is *exclusively* connected with how observed associations *have been* recorded. These features are associated only with labeled associations. For example, \mathbf{X}_k may contain information about the age and extent of research conducted on labeled associations. In our case, this set may contain the total number of scientific publications available at a given time that mention the observed association at least once. This metric serves as a clear proxy for observational bias in mammal-virus associations, as well-studied associations are usually linked to thousands of publications, while newer or less-studied ones may appear in only a few papers.

Additionally, in the following pages we will often assume that we are not given with a complete picture of our database configuration. Instead, we are provided with only a *partial* view of it. In particular, given a set of associations $\Lambda \subset \Omega$, it is possible to define a new structure that encodes our partial information about the database configuration, that is:

$$\mathbb{K}_\Lambda = (S_\Lambda, \mathbf{X}_{k,\Lambda}) = (S|_{\Omega \setminus \Lambda}, \mathbf{X}_k|_{\mathcal{S}(\mathbb{K}) \setminus \Lambda}). \quad (\text{SM7})$$

We will refer to \mathbb{K}_Λ as *partial configuration* or *masked configuration*, subjected to the *mask* Λ . Specifically, \mathbb{K}_Λ possesses information about the labels of all associations in $\Omega \setminus \Lambda$, while associations in Λ are assumed to have both unknown class and unknown label. Also, we define $\mathcal{S}(\mathbb{K}_\Lambda) = \{x \in \Omega \setminus \Lambda : S_\Lambda(x) = 1\}$ and $\mathcal{U}(\mathbb{K}_\Lambda) = \{x \in \Omega \setminus \Lambda : S_\Lambda(x) = 0\}$.

In particular, while our knowledge is bounded to a partial configuration \mathbb{K}_Λ of our database, we can interpret also the label $s \in \{0, 1\}$ for associations $x \in \Lambda$ as a random variable. Specifically, the expression $\Pr(s|\mathbb{K}_\Lambda, x \in \Lambda, x)$ represents the probability of association $x \in \Lambda$ of having label $S(x) = s$, given knowledge of the masked configuration \mathbb{K}_Λ .

Dynamic propensity score

Following the language of classic PU learning [31], we introduce the concept of a *dynamic propensity score*. Consider the probability of an association $x \in \Lambda$ to be labeled, given the information in \mathbb{K}_Λ :

$$\begin{aligned} \Pr(s = 1 \mid \mathbb{K}_\Lambda, x \in \Lambda, x) &= \Pr(s = 1, y = 1 \mid \mathbb{K}_\Lambda, x \in \Lambda, x) = \\ &= \Pr(y = 1 \mid \mathbb{K}_\Lambda, x \in \Lambda, x) \Pr(s = 1 \mid \mathbb{K}_\Lambda, y = 1, x \in \Lambda, x). \end{aligned} \quad (\text{SM8})$$

Here, to obtain the first equivalence, we made use of Supplementary Equation SM6.

We can split the last line of Supplementary Equation SM8 into two parts and analyze the left and right factors separately.

1. The first factor represents the probability of association $x \in \Lambda$ being positive, given the information in \mathbb{K}_Λ . In the context of DPU learning, we assume that this probability can be evaluated if conditioned to some unknown parameters θ that define our statistical model. We can write this as:

$$g_\theta(x; \mathbb{K}_\Lambda) := \Pr(y = 1 \mid \mathbb{K}_\Lambda, x \in \Lambda, x, \theta), \quad (\text{SM9})$$

with g_θ being referred to as the *classifier*. In particular, we notice that $g_\theta(x; \mathbb{K}_\Lambda)$ is well-defined only for associations $x \in \Lambda$.

2. The second factor represents the probability of association $x \in \Lambda$ being labeled, assuming the class it belongs to is positive and given the information in \mathbb{K}_Λ . We will refer to this probability as the *dynamic propensity score* for the association x conditioned to the partial configuration \mathbf{K}_Λ , i.e.,

$$e(x; \mathbb{K}_\Lambda) := \Pr(s = 1 \mid \mathbb{K}_\Lambda, x \in \Lambda, y = 1, x). \quad (\text{SM10})$$

We notice that, contrary to the traditional propensity score in static PU scenarios [31], the dynamic propensity score depends also on the labels of associations in our (masked) database. In particular, $e(x; \mathbb{K}_{\{x\}})$ corresponds to the probability of association x being labeled, assuming x is positive and given the information of the labels of all the other associations $x \in \Omega \setminus \{x\}$. Informally, we can think of $e(x; \mathbb{K}_{\{x\}})$ as answering the question: ‘*What is the probability that, if x were positive, it would have already been observed, given the labels of all the other associations in our database?*’.

Finally, when the precise form of the dynamic propensity score is not known, we can assume it is described by a function parameterized by some parameters ϕ :

$$e_\phi(x; \mathbb{K}_\Lambda) := \Pr(s = 1 \mid \mathbb{K}_\Lambda, x \in \Lambda, y = 1, x, \phi), \quad (\text{SM11})$$

with e_ϕ being referred to as the *propensity score model*. We notice that also this function is well-defined only for associations $x \in \Lambda$.

We will explore a more detailed functional form of g_θ and e_ϕ in the next Subsection (see, in this regard, Subsection 2.3).

With this in mind, we can now ask if it is possible to recover the probability of an instance x to belong to the positive class, given complete knowledge \mathbb{K} over the database configuration at a given time, by relying solely on the probabilities defined in Supplementary Equations SM9 and SM10. The answer is positive, in fact:

$$\begin{aligned} \Pr(y = 1 \mid \mathbb{K}, x, \theta, \phi) &= S(x) + (1 - S(x)) \Pr(y = 1 \mid \mathbb{K}_{\{x\}}, s = 0, x, \theta, \phi) = \\ &= S(x) + (1 - S(x)) \frac{\Pr(s = 0 \mid \mathbb{K}_{\{x\}}, y = 1, x, \phi) \Pr(y = 1 \mid \mathbb{K}_{\{x\}}, x, \theta)}{\Pr(s = 0 \mid \mathbb{K}_{\{x\}}, x, \theta, \phi)} = \\ &= S(x) + (1 - S(x)) \frac{1 - e_\phi(x; \mathbb{K}_{\{x\}})}{1 - e_\phi(x; \mathbb{K}_{\{x\}}) g_\theta(x; \mathbb{K}_{\{x\}})} g_\theta(x; \mathbb{K}_{\{x\}}). \end{aligned} \quad (\text{SM12})$$

In particular, we have used $\mathbb{K} = \mathbb{K}_{\{x\}} \wedge s = 0 \wedge x$, for all x such that $S(x) = 0$, and the fact that labeled instances always belongs to the positive class, as expressed by Supplementary Equation SM6.

This equation provides an operational approach to evaluate the probability that an instance x belongs to the positive class in a DPU scenario, given the configuration of our database at a certain time. It assesses this probability by combining information about the label of x with information about the labels of all other instances in $\Omega \setminus \{x\}$.

If x is labeled in \mathbb{K} , then the class of x is positive as direct consequence of Supplementary Equation SM6. If x is unlabeled in \mathbb{K} , then the class of x must be determined by combining both the propensity score of x and the prediction of the classifier model. A higher classifier output indicates a higher probability of x being positive. Conversely, a higher propensity score suggests a lower probability of x being positive, as x would have already been likely labeled otherwise. Supplementary Equation SM12 integrates both the ground-truth and labeling mechanisms at the same time, and it represents the most complete method for predicting missing links in a DPU scenario during the inference phase.

The assumption of graph-based models

In this Subsection, we focus on how we can model the classifier described in Supplementary Equation SM9. Our first assumption is that this model depends only on the features contained in \mathbf{X}_g , as these are the sole predictive features relevant for distinguishing true from false associations. As emphasized in the previous sections, features in \mathbf{X}_e and \mathbf{X}_k relate primarily to the labeling mechanism rather than to the underlying distribution of true associations. Excluding these features helps to avoid the potentially confounding influences of the labeling process when predicting the true classes of associations.

Furthermore, we assume the model is graph-based. This means that:

$$g_\theta(x; \mathbb{K}_\Lambda) \equiv g_\theta(x; G(\mathbb{K}_\Lambda); \mathbf{X}_g). \quad (\text{SM13})$$

Here, $G(\mathbb{K}_\Lambda) = (\mathcal{V}, E = \mathcal{S}(\mathbb{K}_\Lambda))$ is the graph obtained by representing labeled associations in \mathbb{K}_Λ as edges between the entities in \mathcal{V} .

Similarly, we can assume the propensity score model is graph-based:

$$e_\phi(x; \mathbb{K}_\Lambda) \equiv e_\phi(x; G(\mathbb{K}_\Lambda); \mathbf{X}_e, \mathbf{X}_k). \quad (\text{SM14})$$

However, it is important to note that, although the graph $G(\mathbb{K}_\Lambda)$ is constructed from $\mathcal{S}(\mathbb{K}_\Lambda)$, this represents only a portion of the whole information in S_Λ . In fact, while S_Λ partition Ω in three disjoint sets:

1. $\mathcal{S}(\mathbb{K}_\Lambda) = \{x \in \Omega \setminus \Lambda | S(x) = 1\}$ the set of known labeled associations,
2. $\mathcal{U}(\mathbb{K}_\Lambda) = \{x \in \Omega \setminus \Lambda | S(x) = 0\}$ the set of known unlabeled associations,
3. Λ the set of associations with unknown (i.e., masked) labels;

the graph $G(\mathbb{K}_\Lambda)$ distinguishes only two sets:

1. $E = \mathcal{S}(\mathbb{K}_\Lambda)$ the set of known labeled associations,
2. $\Omega \setminus E = \mathcal{U}(\mathbb{K}_\Lambda) \cup \Lambda$, which includes both known unlabeled associations and associations with unknown (masked) labels.

Thus, the assumption underlying Supplementary Equations SM13 and SM14 is that g_θ and e_ϕ practically treat unlabeled associations and those with unknown label as interchangeable. In reality, this assumption does not generally hold true, as having access to which associations are unlabeled and which are masked can affect predictions. For instance, an association with an unknown label in a region which is known to be dominated by unlabeled associations is less likely to be predicted positive than one in a region dominated by masked labels.

However, if masks are always *sufficiently small and sparse* in the database, the impact of treating masked associations as unlabeled becomes negligible. In particular, while this graph-based assumption does not alter the DPU formalism, it simplifies the problem and allows us to leverage well-established machine learning methods on simple graphs to model the classifier and propensity score model. Since in our work, masks Λ are constructed to be small and sparse in the database (see Subsection 3.2), relying on this approximation is justified and expected to yield reasonable predictions.

2.4 Bayesian DPU model

In this Subsection, we tackle the challenge of estimating probability distributions for propensity scores and positive-class probabilities for each potential mammal-virus association using a Bayesian approach grounded in simple but effective assumptions. By integrating prior knowledge with observational data, the model produces informed estimates—complete with uncertainty quantification—that offer a preliminary solution to our DPU problem.

Formalism of the DPU Bayesian model

For each viral family v , we define a Bayesian model composed of two components: one that generates the probability of an association being positive (i.e., truly existing), and another that estimates the corresponding propensity scores. Let Φ_v and Θ_v denote the parameters of the positive-class probability and propensity score components, respectively, for mammal associations with viral family v . Given a complete configuration \mathbb{K} of the DPU database, we assume that parameters associated to different families are marginally independent, that is:

$$\Pr(\{(\Phi_v, \Theta_v)\}_{v \in V} | \mathbb{K}) = \prod_{v \in V} \Pr(\Phi_v, \Theta_v | \mathbb{K}). \quad (\text{SM15})$$

This assumption implies that the estimations of Φ_v and Θ_v , based on \mathbb{K} , can be performed independently for each viral family $v \in V$.

Also, we assume that the propensity scores associated with viral family v depend, conditioned to the parameters Φ_v , only on the labels of associations related with other viral families $v' \in V \setminus \{v\}$, i.e.,

$$\begin{aligned} e_{\Phi_v}(m, v; \mathbb{K}_{\{(m,v)\}}) &:= \Pr(S(m, v) | \mathbb{K}_{\{(m,v)\}}, m, v, Y(m, v) = 1, \Phi_v) = \\ &= \Pr(S(m, v) | \mathbb{K}_{M \times \{v\}}, m, v, Y(m, v) = 1, \Phi_v) = \\ &=: e_{\Phi_v}(m, v; \mathbb{K}_{M \times \{v\}}). \end{aligned} \quad (\text{SM16})$$

Similarly, the positive-class probability for associations with viral family v do not depend, conditioned to the parameters Θ_v , on the labels of other associations, i.e.,

$$\Pr(Y(m, v) | \mathbb{K}_{\{(m,v)\}}, m, v, \Theta_v) = \Pr(Y(m, v) | m, v, \Theta_v). \quad (\text{SM17})$$

Thus, once the conditional dependencies between the parameters of the Bayesian model and their priors are determined, it is possible to estimate the posterior distribution of the parameters associated with viral family v using:

$$\begin{aligned} \Pr(\Phi_v, \Theta_v | \mathbb{K}) &= \Pr(\Phi_v, \Theta_v | \mathbb{K}_{\{(m,v)\}}, m, v, S(m, v)) \propto \\ &\propto \Pr(S(m, v) | \mathbb{K}_{\{(m,v)\}}, \Phi_v, \Theta_v, m, v) \Pr(\Phi_v, \Theta_v | \mathbb{K}_{\{(m,v)\}}). \end{aligned} \quad (\text{SM18})$$

If, we assume that labels $S(m, v)$ are Bernoulli variables,

$$\Pr(S(m, v) | \mathbb{K}_{\{(m,v)\}}, \Phi_v, \Theta_v, m, v) = p_{m,v}^{S(m,v)} (1 - p_{m,v})^{1-S(m,v)}, \quad (\text{SM19})$$

with

$$\begin{aligned} p_{m,v} &= \Pr(S(m, v) = 1 | \mathbb{K}_{\{(m,v)\}}, \Phi_v, \Theta_v, m, v) = \\ &= e_{\Phi_v}(m, v; \mathbb{K}_{M \times \{v\}}) \Pr(Y(m, v) = 1 | m, v, \Theta_v) \end{aligned} \quad (\text{SM20})$$

being the probability to observe association $(m, v) \in M \times V$, it is then possible to repetitively use Supplementary Equation SM18 to recover the important relation:

$$\Pr(\Phi_v, \Theta_v | \mathbb{K}) \propto \Pr(\Phi_v, \Theta_v) \prod_{m \in M} p_{m,v}^{S(m,v)} (1 - p_{m,v})^{1-S(m,v)}. \quad (\text{SM21})$$

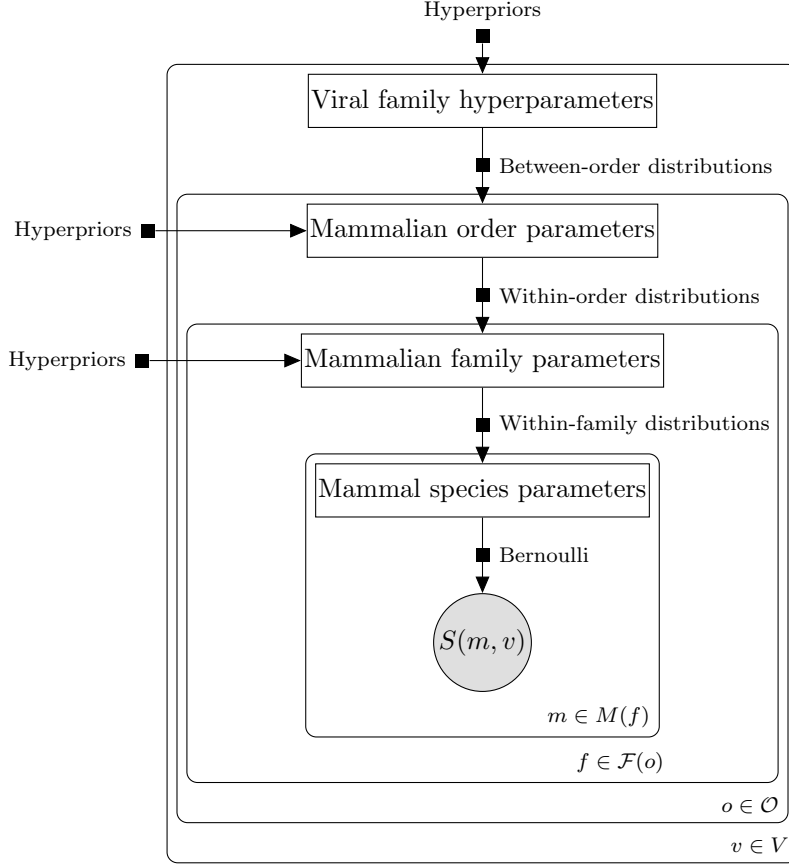
Here, $\Pr(\Phi_v, \Theta_v)$ represents the prior probability associated with the parameters Φ_v and Θ_v of the Bayesian model. It is important to note that deriving Supplementary Equation SM21 would not have been possible if the probabilities $p_{m,v}$, conditioned to parameters Φ_v and Θ_v , had depended on other labeled mammal associations with v . Supplementary Equation SM21 is obtained by repetitively masking the labels of mammal associations with v . If $p_{m,v}$ had been dependent on this masking, calculating the posterior of Φ_v and Θ_v would have required accounting all possible ways of progressively masking the configuration \mathbb{K} , since the masking order would have affected the $p_{m,v}$. This however would have made the calculation of the posterior of the model's parameters infeasible.

Expression (SM21) represents a structured way to estimate the posteriors of the parameters of the propensity score functions and of the class priors for the mammal-virus associations, given observed associations with viral families. However, having to deal with a multi-dimensional problem, the adoption of Markov Chain Monte Carlo (MCMC) methods become indispensable. In this regard, we decided to evaluate the expected values for the propensity scores and the posterior probabilities for the parameters Φ_v and Θ_v using JAGS [32], interfaced with Python [33] through the package pyjags. Sampling was performed on 8 parallel chains with a burn-in of 3000 samples and a thinning interval of 10. The chains were stopped using the R-hat convergence and Effective Sample Size (ESS) efficiency tests [34]; specifically when R-hat values were less than 1.01 and ESS values were greater than 1000 for all parameters and hyperparameters of the model. Calculations were performed using HPC TeraStat2 [35].

Architecture of the Bayesian DPU model

The Bayesian model has an hierarchical structure that reflects mammalian taxonomy. Parameters of mammalian families within an order are drawn from a shared *within-order* distribution, determined by order-level parameters. Each order, in turn, follows a *between-order* distribution defined by hyperparameters specific to each viral family v . To ensure consistency across viral families, we use a shared model architecture for all $v \in V$, with hyperparameters of the same type governed by same hyperpriors. See Supplementary Figures SM2 and SM3 for schematic representations of the model.

The Bayesian model considers 5330 mammal species $m \in M$, 33 viral families $v \in V$, 29 mammalian orders $o \in \mathcal{O}$, and 156 mammalian families $f \in \mathcal{F}$. Also, $\mathcal{F}(o) = \{f \in \mathcal{F} | f \in o\} \subset \mathcal{F}$ is the set of all mammalian families within order $o \in \mathcal{O}$. Similarly, $M(f) = \{m \in M | m \in f\} \subset M$ is the set of all mammal species within mammalian family $f \in \mathcal{F}$. Finally, $f(m) \in \mathcal{F}$ denotes the mammalian family containing mammal species m , while $o(f) \in \mathcal{O}$ denotes the mammalian order containing mammalian family f . For simplicity, we indicate with $o(m) = o(f(m))$ the mammalian order of mammal species m .



Supplementary Figure SM2 | Representation of a hierarchical Bayesian model based on mammalian taxonomy. The model estimates the probability of observed associations $S(m, v)$ between mammal species m and viral families v using a hierarchy of parameters: species-level parameters are drawn from family-level distributions, family-level parameters are drawn from order-level distributions, and order-level parameters are drawn from distributions defined at the viral family level. Each level is governed by its own set of hyperpriors. This hierarchical structure enables information sharing across taxonomic levels, while also allowing each level the flexibility to differ sufficiently from the others. The likelihood is modeled as a Bernoulli distribution at the species level.

Positive-class probabilities. We assume that mammal species belonging to the same mammalian family share a common probability to have a true association with a given viral family. Mathematically:

$$\Pr(Y(m, v) = 1 | m, v, \Theta_v) = \pi_{f(m), v}. \quad (\text{SM22})$$

Here, $\pi_{f, v}$ denotes the parameter in Θ_v that define the positive-class probability for associations between mammal species $m \in M$, belonging to mammalian family f , and viral family $v \in V$. The distributions of these parameters are defined by within-order models specific to the mammalian orders o to which mammalian families f belong:

$$\pi_{f, v} | \mu_{o(f), v}^{(\pi)}, \nu_{o(f), v}^{(\pi)} \sim \text{Beta} \left(\alpha(\mu_{o(f), v}^{(\pi)}, \nu_{o(f), v}^{(\pi)}), \beta(\mu_{o(f), v}^{(\pi)}, \nu_{o(f), v}^{(\pi)}) \right). \quad (\text{SM23})$$

Here, $\mu_{o(f), v}^{(\pi)}$ and $\nu_{o(f), v}^{(\pi)}$ represent the mean and variability scaling factor, respectively, of the Beta distribution associated with mammalian order $o(f) \in \mathcal{O}$. These parameters uniquely determine the shape parameters $\alpha(\mu, \nu)$ and $\beta(\mu, \nu)$ of the distribution, as described in Supplementary Notes, Section 1. We adopt the Beta distribution due to its flexibility and suitability for modeling probabilities constrained to the $[0, 1]$.

We assume that parameters $\mu_{o, v}^{(\pi)} \in (0, 1)$ are defined by a between-orders distribution specific to the viral family v considered:

$$\mu_{o, v}^{(\pi)} | \mu_v^{(\pi)}, \nu_v^{(\pi)} \sim \text{Beta} \left(\alpha(\mu_v^{(\pi)}, \nu_v^{(\pi)}), \beta(\mu_v^{(\pi)}, \nu_v^{(\pi)}) \right) \quad \text{T}(0.001, 0.999). \quad (\text{SM24})$$

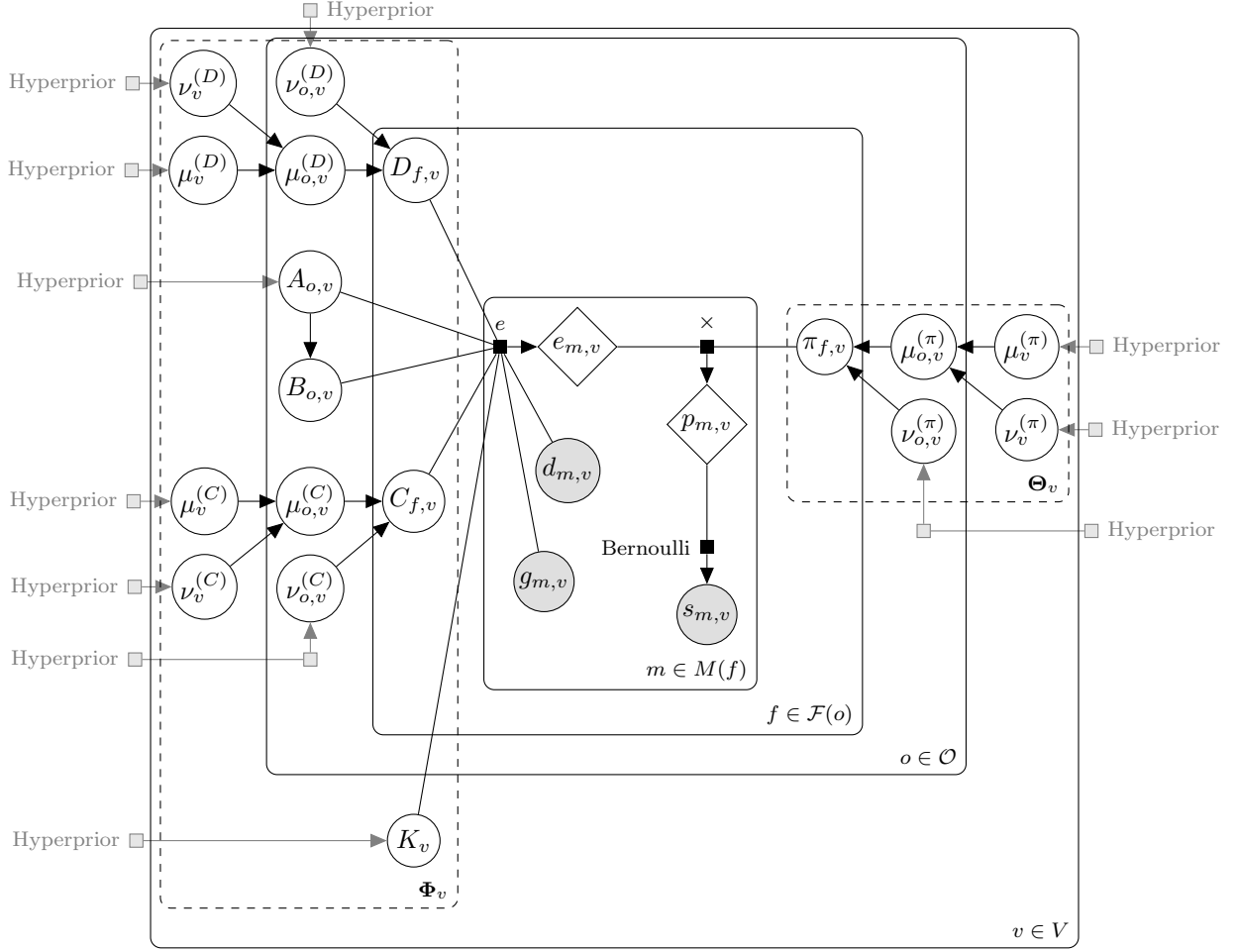
Here, $\text{T}(0.001, 0.999)$ indicates truncation of the Beta distribution to avoid boundary values.

Finally, the parameters $\nu_{o, v}^{(\pi)}$, as well as $\mu_v^{(\pi)}$ and $\nu_v^{(\pi)}$, are the hyperparameter of the model associated to viral family v and govern the overall behavior of each viral family.

Propensity scores. We model propensity scores as:

$$e_{\Phi_v}(m, v; \mathbb{K}_{M \times \{v\}}) = e(d_m(\mathbb{K}_{M \times \{v\}}), g_m(\mathbb{K}_{M \times \{v\}}); \phi_{f(m), v}). \quad (\text{SM25})$$

Here:



Supplementary Figure SM3 | Representation of the Bayesian model as a directed acyclic graph. The model estimates the probability of a mammal-virus association $s_{m,v}$ as a Bernoulli variable with success probability $p_{m,v} = e_{m,v} \cdot \pi_{f,v}$, where $e_{m,v}$ is a propensity score capturing geographical, phylogenetic, and human-driven labeling biases for the mammal-virus pair (m, v) , and $\pi_{f,v}$ reflects the underlying likelihood of association between mammalian family f and viral family v . White circles denote unobserved variables, gray circles represent observed variables, and white diamonds indicate deterministic but unobserved variables.

- $\phi_{f(m),v}$ is a set of parameters within Φ_v , that depends from the mammalian family $f(m)$ and from the viral family v ;
- $d_m(\mathbb{K}_{M \times \{v\}})$ is the degree of the mammalian species m in the masked database $\mathbb{K}_{M \times \{v\}}$;
- $g_m(\mathbb{K}_{M \times \{v\}})$ is a proxy measure of mean geographic exposure of species m to field studies related to viruses in viral families different from v .

Specifically:

$$d_m(\mathbb{K}_{M \times \{v\}}) = \begin{cases} d_m(\mathbb{K}) - 1 & \text{if } (m, v) \in \mathcal{S}(\mathbb{K}), \\ d_m(\mathbb{K}) & \text{otherwise,} \end{cases} \quad (\text{SM26})$$

where $d_m(\mathbb{K})$ is the degree of mammalian species m in \mathbb{K} , that is the number of its observed associations to viral families in the database.

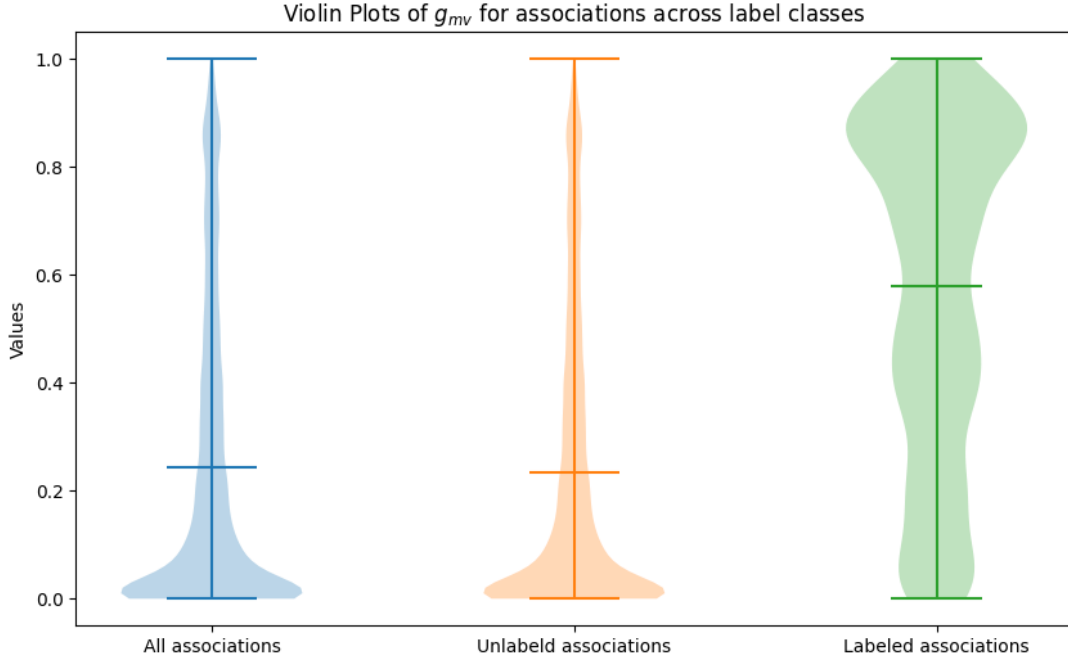
On the other hand, the proxy measure $g_m(\mathbb{K}_{M \times \{v\}})$ is defined as:

$$g_m(\mathbb{K}_{M \times \{v\}}) = \frac{1}{|V| - 1} \sum_{v' \in V \setminus \{v\}} \left(\max_{m' \in M \setminus \{m\}} g(m, m', v') \right), \quad (\text{SM27})$$

with:

$$g(m, m', v') := \begin{cases} \frac{\text{overlap}(m', m)}{r_{m'}} & \text{if } (m', v') \in \mathcal{S}(\mathbb{K}), \\ 0 & \text{otherwise.} \end{cases} \quad (\text{SM28})$$

Here, $\text{overlap}(m', m)/r_{m'}$ represents the geographic overlap between m and m' normalized by the range $r_{m'}$ of species m' . Thus, $\max_{m' \in M \setminus \{m\}} g(m, m', v')$ is maximal if species m is ubiquitous in the range of at least one species m' known to be associated to viral family v' . By averaging over all viral families $v \in V \setminus \{v'\}$, this measure allows us



Supplementary Figure SM4 | Predictive role of $g_{m,v}$ for labeled associations. Violin plots of $g_{m,v}$ across all possible mammal–virus pairs, comparing distributions for labeled and unlabeled associations in the VIRION database. Higher values of $g_m(\mathbb{K}_{M \times \{v\}})$ are associated with an increased likelihood of an observed (labeled) association.

to quantify the mean geographic exposure of species m to field studies, enabling us to assess the potential impact of geographic bias on the likelihood of species m being studied. Also, $g_m(\mathbb{K}_{M \times \{v\}})$ can be shown to be an important predictor for determining labeled associations, as it is possible to notice by looking in Supplementary Figure SM4.

This approach enables the generation of tailored propensity scores that are not only sensitive to specific combinations of viral families and mammalian taxonomic groups, but that also consider how well-studied species m is—both in terms of its associations with other viral families and the intensity of research conducted within its geographic range. As a result, this framework can capture both virus–taxon-specific biases (including biological and self-reinforcing biases), as well as geographic and human-centric biases. Notably, a species’ degree often serves as a proxy for scientific and economic attention it receives.

Thus, we propose the following functional form for the propensity scores:

$$e(d, g; \phi_{f,v}) = \frac{1 + K_v \cdot g}{1 + K_v \cdot g \cdot e_0(d, \phi_{f,v})} \cdot e_0(d; \phi_{f,v}), \quad (\text{SM29})$$

with:

$$e_0(d; \phi_{f,v}) = C_{f,v} + \frac{\sigma\left(\frac{d - A_{f,v}}{B_{f,v}}\right) - \sigma\left(\frac{-A_{f,v}}{B_{f,v}}\right)}{1 - \sigma\left(\frac{-A_{f,v}}{B_{f,v}}\right)} \cdot D_{f,v}(1 - C_{f,v}), \quad (\text{SM30})$$

and with $\phi_{f,v} \equiv (K_v; A_{f,v}, B_{f,v}, C_{f,v}, D_{f,v})$ and $\sigma(x) = 1/(1 + e^{-x})$ being the sigmoid function. Specifically, $C_{f,v} \in [0, 1]$ and $D_{f,v} \in [0, 1]$, while $K_v, A_{f,v} \geq 0$ and $B_{f,v} > 0$.

The baseline function $e_0(d; \phi_{f,v})$ models the dependence of the propensity scores from d and represents the continuous equivalent of a two-level system. For poorly-studied species (with $d \approx 0$), we have:

$$e_0(d; \phi_{f,v}) \approx C_{f,v},$$

while well-studied species (with high d , i.e., $d > A_{f,v} + 4B_{f,v}$), we have:

$$e_0(d; \phi_{f,v}) \approx D_{f,v}(1 - C_{f,v}) + C_{f,v} \geq C_{f,v}.$$

Species with intermediate degrees are modeled through a logistic interpolation, with flex in $A_{f,v}$ and steepness governed by $B_{f,v}$.

On the other hand, Supplementary Equation SM29 introduces a correction term for accounting geographic sampling biases. In particular, for species with low $e_0(d; \phi_{f,v})$, the propensity scores can be approximated as:

$$e(d, g; \phi_{f,v}) \approx (1 + K_v g) e_0(d; \phi_{f,v}) \geq e_0(d; \phi_{f,v}),$$

while for species with high $e_0(d; \phi_{f,v})$, the geographic exposure has little impact, and the model simplifies to:

$$e(d, g; \phi_{f,v}) \approx e_0(d; \phi_{f,v}).$$

Supplementary Table SM3 | Hyperparameters and their hyperprior distributions in the DPU Bayesian model. The table summarizes the hyperparameters common to all viral families ($v \in V$), their corresponding hyperprior distributions, along with the parameters of these hyperpriors, and their associated (0.02, 0.98) quantile ranges (QR). The hyperpriors are chosen to discourage extreme parameter values while allowing realistic variability in the viral spread and labeling bias across mammalian species.

Hyperparam	Hyperprior	Hyperprior's parameters	(0.02, 0.98)-QR
$\mu_v^{(\pi)}$	Beta	$\alpha = 2.294, \beta = 2.294$	(0.101, 0.899)
$\nu_v^{(\pi)}$	Beta	$\alpha = 4.409, \beta = 3.269$	(0.225, 0.883)
$\nu_{o,v}^{(\pi)}$	Beta	$\alpha = 2.833, \beta = 2.833$	(0.128, 0.872)
$\mu_v^{(C)}$	Beta	$\alpha = 2.303, \beta = 109$	(0.003, 0.056)
$\nu_v^{(C)}$	Beta	$\alpha = 2.437, \beta = 1.166$	(0.182, 0.985)
$\nu_{o,v}^{(C)}$	Beta	$\alpha = 1.967, \beta = 1.108$	(0.127, 0.985)
$\mu_v^{(D)}$	Beta	$\alpha = 2.305, \beta = 2.823$	(0.086, 0.853)
$\nu_v^{(D)}$	Beta	$\alpha = 2.437, \beta = 1.166$	(0.182, 0.985)
$\nu_{o,v}^{(D)}$	Beta	$\alpha = 1.967, \beta = 1.108$	(0.127, 0.985)
$A_{o,v}$	$15 \times \text{Beta}$	$\alpha = 2, \beta = 2$	(1.261, 13.739)
$B_{o,v} A_{o,v}$	Normal	see text	variable
K_v	Exponential	$\lambda = 0.921$	(0.022, 4.248)

Similarly to what done with the parameters $\pi_{f,v}$, we also assume that the parameters $C_{f,v} \in [0, 1]$ and $D_{f,v} \in [0, 1]$ are generated by a hierarchical model based on mammalian taxonomy.

Specifically, the family-level parameters, $C_{f,v}$ and $D_{f,v}$, are drawn from the following within-order distributions:

$$C_{f,v}|\mu_{o(f),v}^{(C)}, \nu_{o(f),v}^{(C)} \sim \text{Beta}\left(\alpha(\mu_{o(f),v}^{(C)}, \nu_{o(f),v}^{(C)}), \beta(\mu_{o(f),v}^{(C)}, \nu_{o(f),v}^{(C)})\right), \quad (\text{SM31})$$

$$D_{f,v}|\mu_{o(f),v}^{(D)}, \nu_{o(f),v}^{(D)} \sim \text{Beta}\left(\alpha(\mu_{o(f),v}^{(D)}, \nu_{o(f),v}^{(D)}), \beta(\mu_{o(f),v}^{(D)}, \nu_{o(f),v}^{(D)})\right), \quad (\text{SM32})$$

while parameters $\mu_{o(f),v}^{(C)} \in (0, 1)$ and $\mu_{o(f),v}^{(D)} \in (0, 1)$ are drawn by the following between-order distributions:

$$\mu_{o,v}^{(C)}|\mu_v^{(C)}, \nu_v^{(C)} \sim \text{Beta}\left(\alpha(\mu_v^{(C)}, \nu_v^{(C)}), \beta(\mu_v^{(C)}, \nu_v^{(C)})\right) \quad \text{T}(0.0001, 0.999), \quad (\text{SM33})$$

$$\mu_{o,v}^{(D)}|\mu_v^{(D)}, \nu_v^{(D)} \sim \text{Beta}\left(\alpha(\mu_v^{(D)}, \nu_v^{(D)}), \beta(\mu_v^{(D)}, \nu_v^{(D)})\right) \quad \text{T}(0.001, 0.999), \quad (\text{SM34})$$

with $\mu_v^{(C)}$, $\nu_v^{(C)}$ and $\mu_v^{(D)}$, $\nu_v^{(D)}$ being hyperparameters specific to the viral family v considered.

Conversely, the parameters $A_{f,v}$ and $B_{f,v}$ are directly defined at the order-level:

$$A_{f,v} \equiv A_{o(f),v}, \quad B_{f,v} \equiv B_{o(f),v}, \quad (\text{SM35})$$

and are assumed to be hyperparameters specific to the viral family v . Finally, K_v completes the last hyperparameter of the propensity score model related to viral family $v \in V$.

Bayesian model's hyperpriors

Hyperpriors are chosen to discourage extreme values and ensure realistic variability in the spread of viral families across mammalian orders. In Supplementary Table SM3 is a summary of the hyperparameters and their hyperpriors. We remember hyperpriors are common to all viral families $v \in V$.

- **Hyperparameters μ and ν .** All the means μ and the variability scaling factors ν follow Beta distributions as they are constrained to be in the interval $(0, 1)$.

For the variability scaling factors ν , we impose hyperpriors to reflect the assumption that within-order variability is generally lower than between-order variability. Furthermore, we assume that the variability in the probabilities of association within a taxonomic unit is lower than the variability in propensity scores within the same unit. The priors of $\nu_v^{(\pi)}$ and $\nu_{o,v}^{(\pi)}$ are chosen so that there is a 50% probability that all $\nu_v^{(\pi)}$ values fall within the range (0.2, 0.9), and that, for a specific order o , all $\nu_{o,v}^{(\pi)}$ values fall within the range (0.1, 0.9). On the other hand, the priors of $\nu_v^{(C)}$ and $\nu_{o,v}^{(C)}$, as well as $\nu_v^{(D)}$ and $\nu_{o,v}^{(D)}$, are chosen so that there is a 50% probability that all $\nu_v^{(C)}$ and all $\nu_v^{(D)}$ values fall within the range (0.2, 0.99), and that, for a specific order o , all $\nu_v^{(C)}$ and all $\nu_v^{(D)}$ values fall within the range (0.1, 0.99). This reflects our belief that the disparities in research focus and resources allocated to studying various taxonomic groups are more pronounced than the actual biological differences in

their susceptibility to viruses. Supplementary Figure SM5 illustrates how different priors for ν affect the shape of the resulting distribution.

For the mean hyperparameters, the prior of $\mu_v^{(\pi)}$ is chosen such that there is a 50% chance that, a priori, all $\mu_v^{(\pi)}$ values fall inside the range (0.075, 0.925). The prior of $\mu_v^{(C)}$ is chosen to ensure that, a priori, there is a 50% probability that no mammalian order has a family f with 50% probability of having $C_{v,f} > 0.5$ or a family f with 50% probability of having $C_{v,f} < 1/(3904 \times 33)$, while the prior of $\mu_v^{(D)}$ to ensure that, a priori, there is a 50% probability that at least one mammalian order has a family f with 50% probability of having $D_{v,f} > 0.5$ and at least one mammalian order has a family f with 50% probability of having $D_{v,f} < 0.25$. Here, 3904 is the number of mammal species in M with no observed associations to viruses, while 33 is $|V|$.

- **Hyperparameters $A_{o,v}$ and $B_{o,v}$.** To ensure that propensity scores saturate at high degrees, we propose that $A_{o,v}$ should be strictly lower than 15. Also, it is necessary to constrain the maximum value that $B_{o,v}$ to ensure that e_0 does not become too shallow and fails to saturate when the mammalian degree $d \sim d_{\max} = 22$. Considering that $\sigma(5) \approx 0.99$, we require $B_{o,v} < (d_{\max} - A)/5$, with $d_{\max} = 22$. Similarly, we require that $B_{o,v} > 2/3$, since otherwise the sigmoid function would become excessively steep, transitioning from approximately 0.18 to 0.82 within less than a two degree step. If we assume that:

$$B_{o,v}|A_{o,v} \sim \text{Normal}\left(\mu^{(B)}(A_{o,v}), \sigma^{(B)}(A_{o,v})\right), \quad (\text{SM36})$$

and that 98% of the probability mass of this distribution is within the above stated bounds for $A_{o,v} = 0$ and $A_{o,v} = 15$, we find:

$$\mu^{(B)}(0) = 1.033, \quad \sigma^{(B)}(0) = 0.158, \quad (\text{SM37})$$

$$\mu^{(B)}(15) = 2.533, \quad \sigma^{(B)}(15) = 0.802. \quad (\text{SM38})$$

Interpolating linearly between them, we recover a general form for the parameters of the prior of $B_{o,v}$ conditioned to the value of $A_{o,v}$:

$$\mu^{(B)}(A_{o,v}) = \mu^{(B)}(0) + (\mu^{(B)}(15) - \mu^{(B)}(0)) \times A_{o,v}/15, \quad (\text{SM39})$$

$$\sigma^{(B)}(A_{o,v}) = \sigma^{(B)}(0) + (\sigma^{(B)}(15) - \sigma^{(B)}(0)) \times A_{o,v}/15. \quad (\text{SM40})$$

- **Hyperparameter K_v .** For the prior of K_v , we use an Exponential distribution with a rate parameter that ensures 99% of the distribution's values fall within the range $[0, 5)$. Using an Exponential distribution also means that, a priori, the most likely value for K_v is 0, representing the most conservative initialization.

2.5 Classifier model

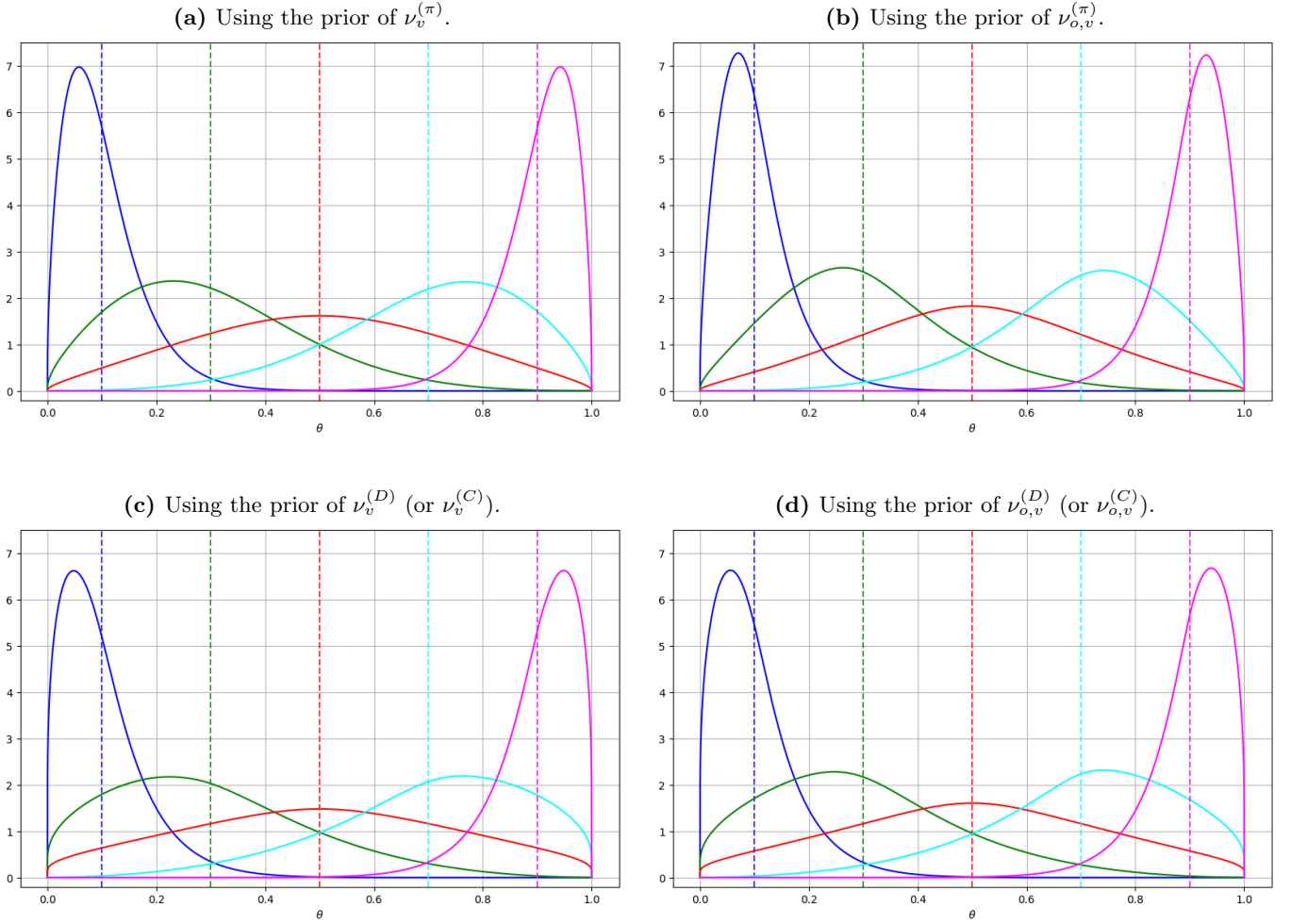
In this Subsection, we analyze the model proposed for the classifier. This model aims to reproduce the probability a mammal-virus association belongs to the positive class given a partial graph of known mammal-virus associations $G = (M \cup V, E \subset M \times V)$ and a set of parameters θ . Mathematically: $g_\theta(x; G) = \Pr(y = 1|G, x, \theta)$, with $x \in M \times V \setminus E$. The model is divided into two distinct and sequential modules:

1. **Virus-Mammal Graph Neural Network (ViM-GNN).** The first module employs a bipartite Graph Neural Network model to generate meaningful node representations for both virus and mammal nodes. These are designed to capture the preferred associations between viruses and mammals by considering known virus-mammal interactions, phylogenetic and geographic relationships between mammal species, as well as mammal-specific and viral-specific features.
2. **Link Prediction Classifier.** The second and final module is a fully connected Dense Neural Network that uses the node representations from the ViM-GNN to predict the presence of a true association between a couple of virus and mammal nodes.

The two modules are trained end-to-end via backpropagation. Finally, we remember that we have defined \mathbf{x}_m being the mammalian feature vector (of dimension 18) associated to mammal nodes $m \in M$ and \mathbf{x}_v being the viral feature vector (of dimension 7) associated to viral nodes $v \in V$. All node features are normalized feature-wise such that each feature has a mean of 0 and a variance of 1 across all nodes.

Virus-Mammal Graph Neural Network

To effectively model the positional context of nodes within the observed multi-graph of virus-mammal interactions, we construct multiple latent adjacency matrices.



Supplementary Figure SM5 | Effect of different hyperpriors on the variability of parameter distributions. Plots of $\Pr(\theta | \mu) = \int_0^1 \Pr(\theta | \mu, \nu) \Pr(\nu) d\nu$, using different priors for $\Pr(\nu)$ and fixed values of $\mu \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$. Vertical dotted lines indicate the corresponding mean values μ .

Given a graph of known mammal-virus associations $G = (M \cup V, E(G) \subset M \times V)$, we define the normalized bipartite adjacency matrices $\mathbf{A}_{VM}^{(G)} : V \times M \rightarrow [0, 1]$ and $\mathbf{A}_{MV}^{(G)} : M \times V \rightarrow [0, 1]$ as:

$$\begin{aligned} \mathbf{A}_{VM}^{(G)}[v, m] &= \begin{cases} \frac{1}{\sqrt{d_v^{(G)} d_m^{(G)}}}, & \text{if } (m, v) \in E(G), \\ 0, & \text{otherwise,} \end{cases} \\ \mathbf{A}_{MV}^{(G)}[m, v] &= \begin{cases} \frac{1}{\sqrt{d_v^{(G)} d_m^{(G)}}}, & \text{if } (m, v) \in E(G), \\ 0, & \text{otherwise.} \end{cases} \end{aligned} \quad (\text{SM41})$$

Here, $d_v^{(G)}$ and $d_m^{(G)}$ represent the degrees of the virus and mammal nodes, v and m , in G .

We adopted symmetric normalization, as introduced by Kipf & Welling (2017) [36], because, compared to random walk normalization, it amplifies signals from low-degree nodes relative to high-degree ones. This choice is particularly important in observed mammal-virus networks, where low-degree nodes—often corresponding to associations with low propensity scores—carry valuable but rare information. By mitigating the dominance of high-degree nodes during message passing, symmetric normalization helps preserve these signals in the aggregation process, improving the model’s ability to detect underrepresented interaction patterns.

Let $\mathbf{M} : M \times M \rightarrow [0, 1]$ denote a mammal-mammal interaction adjacency matrix. Using this, we define the latent adjacency matrix $\mathbf{Q}_{MV} : M \times V \rightarrow [0, 1]$ as:

$$\mathbf{Q}_{MV}^{(\mathbf{M}, G)} = (\mathbf{D}_{\mathbf{MA}_{MV}^{(G)}})^{-1} \mathbf{MA}_{MV}^{(G)}, \quad (\text{SM42})$$

where $\mathbf{D}_{\mathbf{MA}_{MV}^{(G)}} : M \times M \rightarrow \mathbb{R}^+$ is the diagonal matrix of row sums of $\mathbf{MA}_{MV}^{(G)}$.

This latent matrix enables viruses to connect indirectly to mammal species through mammal-mammal relationships

Algorithm 1: Pseudocode for generating $\mathbf{Q}_{MV}^{(M,G)}$

Input : mammal-virus adjacency matrix $\mathbf{A}_{MV}^{(G)}$
mammal-mammal adjacency matrix \mathbf{M}
 $q \in [0, 1]$

Output: sparse mammal-virus adjacency matrix $\mathbf{Q}_{MV}^{(M,G)}$

- 1 $\mathbf{Q}_{MV}^{(M,G)} \leftarrow \text{MatMul}(\mathbf{M}, \mathbf{A}_{MV}^{(G)});$
- 2 $\mathbf{Q}_{MV}^{(M,G)} \leftarrow \text{RndWalkNorm}(\mathbf{Q}_{MV}^{(M,G)});$
- 3 $\theta_q \leftarrow \text{Quantile}_q(\mathbf{Q}_{MV}^{(M,G)});$
- 4 **foreach** $(m, v) \in M \times V$ **do**
- 5 **if** $\mathbf{Q}_{MV}^{(M,G)}[m, v] < \theta_q$ **then**
- 6 $\mathbf{Q}_{MV}^{(M,G)}[m, v] \leftarrow 0;$
- 7 $\mathbf{Q}_{MV}^{(M,G)} \leftarrow \text{RndWalkNorm}(\mathbf{Q}_{MV}^{(M,G)});$
- 8 **return** $\mathbf{Q}_{MV}^{(M,G)}$

with species they are already known to infect, effectively constructing meta-paths between viruses and potential hosts. These meta-paths, akin to those in Graph Transformer Networks [37], are composite pathways that traverse both the observed virus-mammal and the fixed mammal-mammal networks. Their introduction allows us to characterize viral diversity within mammalian groups where direct virus-host associations are sparse or absent by leveraging known interspecific relationships.

However, the resulting matrix $\mathbf{Q}_{MV}^{(M,G)}$ tends to be quite dense, as in most cases, each mammal species becomes connected via the latent matrix to all (or nearly all) viral families. This leads to over-smoothing in GNN models. To mitigate this problem, we prune (i.e., set to zero) all connections below a threshold defined by the q -quantile of the elements in $\mathbf{Q}_{MV}^{(M,G)}$ and re-normalize the resulting matrix via random-walk normalization. The complete procedure is found in the pseudocode reported in 8. Best performance was found using $q = 0.43$.

One might reasonably define the matrix for the reverse direction as:

$$\mathbf{Q}_{VM}^{(M,G)} = (\mathbf{D}_{\mathbf{A}_{VM}^{(G)}\mathbf{M}^T})^{-1} \mathbf{A}_{VM}^{(G)} \mathbf{M}^T, \quad (\text{SM43})$$

but this leads to performance degradation. Instead, we conservatively use:

$$\mathbf{Q}_{VM}^{(M,G)} = \mathbf{D}_{\mathbf{A}_{VM}^{(G)}}^{-1} \mathbf{A}_{VM}^{(G)}. \quad (\text{SM44})$$

leveraging only known associations. Given that each virus is typically linked to multiple mammal hosts, this suffices for signal propagation from viral to mammal nodes.

ViM-GNN iteratively updates node embeddings via message-passing using different latent matrices. At each step, $k \in \{1, \dots, K\}$, given a mammal-mammal interaction network \mathbf{M} , node embeddings updates as follows:

$$\mathbf{K}_S^{(M,k)} = \text{MLP}_{TS}^{(M,k)} \left(\mathbf{H}_S^{(M,k-1)} \right), \quad (\text{SM45})$$

$$\mathbf{K}_{\mathcal{N}(T)}^{(M,k)} = \text{GraphNorm}_{TS}^{(M,k)} \left(\mathbf{Q}_{TS}^{(M,G)} \mathbf{K}_S^{(M,k)} \right), \quad (\text{SM46})$$

$$\mathbf{K}_T^{(M,k)} = \text{SLP}_{TS, \text{t-update}}^{(M,k)} \left(\mathbf{H}_T^{(M,k-1)} \right), \quad (\text{SM47})$$

$$\mathbf{H}_T^{(M,k)} = \text{SLP}_{TS, \text{update}}^{(M,k)} \left(\mathbf{K}_T^{(M,k)} \oplus \mathbf{K}_{\mathcal{N}(T)}^{(M,k)} \right). \quad (\text{SM48})$$

Where:

- $T, S \in \{M, V\}$, represents target and source node sets, i.e., if target nodes are of mammal type, $T = M$ and $S = V$,
- \oplus stands for concatenation,
- MLP and SLP are multi- and single-layer perceptrons with non-linearity applied using PReLU [38],
- initial node embeddings are given by nodes features, i.e., $\mathbf{h}_v^{(M,0)} = \mathbf{x}_v$ and $\mathbf{h}_m^{(M,0)} = \mathbf{x}_m$.

In our case, latent propagation is guided by only two mammal interaction networks $\mathfrak{M} = \{\hat{\mathbf{G}}, \mathbf{P}_\tau\}$. Here, $\hat{\mathbf{G}}$ is the generalized biogeographic adjacency matrix, defined as:

$$\hat{\mathbf{G}} = c\mathbf{G} + (1 - c)\mathbf{G}^2 = \mathbf{G}(c + (1 - c)\mathbf{G}), \quad (\text{SM49})$$

where $c \in [0, 1]$ is a learnable weight in the model, and \mathbf{G}, \mathbf{G}^2 represent the 1-hop and 2-hop biogeographic adjacency matrices, respectively, as defined in Supplementary Equations SM4 and SM5. This formulation allows us to capture both direct and 2-hop interactions, while preserving normalization.

\mathbf{P}_τ is a phylogenetic adjacency matrix, as described in Supplementary Equation SM3, with phylogenetic temperatures $\tau = 0.07 \times t_{\max}$, where $t_{\max} = \max_{m, m'} t(m, m') = 3.16 \times 10^{11}$ yrs is the maximal patristic distance between all couples of mammal species. Although we used a single phylogenetic adjacency matrix in this work, latent mammal interaction networks can be extended to incorporate multiple phylogenetic matrices computed at different temperature parameters, allowing the model to capture phylogenetic relationships across varying evolutionary lengths.

Finally, after $K = 3$ message-passing iterations, the node embeddings produced from different mammal-mammal interaction networks are aggregated into a unified embedding \mathbf{z} using an attention-based mechanism:

$$\mathbf{z}_t = \frac{\sum_{\mathbf{M} \in \mathfrak{M}} \mathbf{w}_T^{(\mathbf{M})} \odot e^{\beta \odot \mathbf{h}_t^{(K, \mathbf{M})}} \odot \mathbf{h}_t^{(K, \mathbf{M})}}{\sum_{\mathbf{M} \in \mathfrak{M}} \mathbf{w}_T^{(\mathbf{M})} \odot e^{\beta \odot \mathbf{h}_t^{(K, \mathbf{M})}}}, \quad (\text{SM50})$$

where \odot denotes element-wise multiplication, $\mathbf{w}_T^{(\mathbf{M})}$ are learnable attention weights satisfying $\sum_{\mathbf{M} \in \mathfrak{M}} \mathbf{w}_T^{(\mathbf{M})} = \mathbf{1}_T$, and β are learnable inverse temperature parameters.

This aggregation scheme is a weighted generalization of the SoftMax Aggregation proposed in [39], allowing a smooth transition between weighted mean, max, and min aggregations. For small β , Supplementary Equation SM50 approximates a weighted mean with weights $\mathbf{w}_T^{(\mathbf{M})}$. For large positive (negative) β , it approximates a max (min) aggregation. The mechanism thus combines attention weights $\mathbf{w}_T^{(\mathbf{M})}$ with a content-sensitive modulation via β , enabling more expressive and adaptive aggregation.

The parameters $\mathbf{w}_T^{(\mathbf{M})}$ are obtained via a feature-wise softmax over a trainable matrix initialized uniformly in $(-2, 2)$, while β is initialized as a constant vector with all entries set to 0.1. Empirically, this approach outperforms both simple weighted averaging and concatenation of embeddings.

In our setup, both for mammalian species and viral family nodes, the dimensionality of the output node embeddings were set to 64 for the first, second, and final layers.

Link Prediction Classifier

The module uses the embeddings generated by the ViM-GNN for mammal and virus nodes to infer the probability of a true association between them. The module is defined as:

$$\text{pred}(m, v) = \text{MLP}(\mathbf{z}_m \oplus \mathbf{z}_v), \quad (\text{SM51})$$

where \oplus denotes the concatenation operator and MLP refers to a multilayer perceptron with a single hidden layer. The activation function of the hidden layer is PRELU, while the output layer uses the sigmoid function to ensure the results fall within the range $[0, 1]$, making them interpretable as probabilities.

The Link Prediction Classifier input has a dimension of 128, as the final embedding dimension for both mammal and virus nodes is 64. The MLP’s hidden layer consists of 64 neurons.

3 Model Training

This section offers supplementary information on the strategy used to train our machine learning models based on the positive and unlabeled mammal–virus associations.

- Subsection 3.1 introduces DPU learning, a general framework for training models in DPU scenarios.
- Subsection 3.2 presents our subgraph sampling scheme, used to construct training, validation, and test sets from the observed graph of mammal-virus associations.
- Subsection 3.3 details our specific training implementation, combining the DPU framework with graph sampling to address our case study.

3.1 Dynamic Positive-Unlabeled learning

Here, we present one generalizable strategy designed to allow the training of unbiased classifiers in scenarios adhering to the DPU framework. We assume access to the exact dynamic propensity scores and to the configuration \mathbb{K} of the database.

Theoretical risk in DPU learning

Given a loss function $\ell : [0, 1] \times 0, 1 \rightarrow \mathbb{R}^+$, we define the theoretical risk as:

$$\begin{aligned} R_{\ell, \mathbb{K}, \mathcal{M}}^{\text{DPU}}[g] := & \frac{1}{|\Omega| \cdot |\mathcal{M}|} \sum_{x \in \mathcal{P}} \sum_{\Lambda \in \mathcal{M}(x)} \ell(g(x, G(\mathbb{K}_\Lambda); \mathbf{X}_g), 1) + \\ & + \frac{1}{|\Omega| \cdot |\mathcal{M}|} \sum_{x \in \mathcal{N}} \sum_{\Lambda \in \mathcal{M}(x)} \ell(g(x, G(\mathbb{K}_\Lambda); \mathbf{X}_g), 0), \end{aligned} \quad (\text{SM52})$$

where $\mathcal{M}(x)$ is a fixed-size of masks containing x , i.e., $\forall x \in \Omega$, $|\mathcal{M}(x)| = |\mathcal{M}|$ and $x \in \Lambda$ for all $\Lambda \in \mathcal{M}(x)$.

This risk in Supplementary Equation SM52 evaluates model g 's ability to infer the correct class of each instance $x \in \Omega$, conditioned on multiple, diverse partial views \mathbb{K}_Λ of the full database. The choice of mask sets $\mathcal{M}(x)$ regulates how well g generalizes across different graph configurations.

When using small-sized masks (e.g., $\Lambda = \{x\}$), g is encouraged to rely on the full structural and attribute information available in \mathbb{K} . In these cases, the model's performance is optimized for situations where full information is available at test time, and the focus is on maximizing classification accuracy under complete observability. In contrast, larger masks force g to make predictions with limited information, by introducing a regularization effect similar to edge dropout in GNNs. This reduces the model's reliance on specific patterns or substructures in the training data and encourages the learning of more generalizable and robust data representations. Moreover, this masking strategy also serves as a form of data augmentation, as by systematically varying between different mask configurations, the model is exposed to more diverse partial views \mathbb{K}_Λ of the same database. This increases the effective variability of the training set, which is especially valuable given the finite and limited space of possible graph configurations in Ω .

Ultimately, the choice of how to construct $\mathcal{M}(x)$ can significantly impact the trade-off between expressiveness and robustness in the learned model. Determining the optimal masking strategy will often depend on the specific task and the characteristics of the downstream application.

Empirical DPU risk estimation

To connect theory to practice, we approximate the theoretical risk using labeled and unlabeled associations in \mathbb{K} . The empirical DPU risk is defined as:

$$\begin{aligned} \hat{R}_{\ell, \mathbb{K}, \mathcal{M}}^{\text{DPU}}[g] := & \frac{1}{|\Omega| \cdot |\mathcal{M}|} \sum_{x \in \mathcal{U}(\mathbb{K})} \sum_{\Lambda \in \mathcal{M}(x)} \ell(g(x; \mathbb{K}_\Lambda; \mathbf{X}_g), 0) + \\ & + \frac{1}{|\Omega| \cdot |\mathcal{M}|} \sum_{x \in \mathcal{S}(\mathbb{K})} \sum_{\Lambda \in \mathcal{M}(x)} \frac{1}{e(x|\mathbb{K}_{\{x\}})} \ell(g(x; \mathbb{K}_\Lambda; \mathbf{X}_g), 1) + \\ & + \frac{1}{|\Omega| \cdot |\mathcal{M}|} \sum_{x \in \mathcal{S}(\mathbb{K})} \sum_{\Lambda \in \mathcal{M}(x)} \left(1 - \frac{1}{e(x|\mathbb{K}_{\{x\}})}\right) \ell(g(x; \mathbb{K}_\Lambda; \mathbf{X}_g), 0), \end{aligned} \quad (\text{SM53})$$

where $e(x|\mathbb{K}_x)$ is the known dynamic propensity score for instance $x \in \Omega$.

This risk generalizes the SAR-PU risk proposed by Bekker et al. (2020) [40] adjusting for the DPU framework. However, unlike the SAR case, the empirical DPU risk (SM53) cannot be demonstrated to be an unbiased estimator of the true risk due to the inapplicability of the law of large numbers arising from the finiteness of the instance space Ω .

The rationale behind the DPU empirical risk is that, for each labeled association $x \in \mathcal{S}(\mathbb{K})$, we expect approximately $1/e(x|\mathbb{K}_x)$ similar true positives in total. This means that $1 - 1/e(x|\mathbb{K}_x)$ of them are likely still unlabeled. To account for this imbalance and the uncertainty in labeling, the risk function incorporates the following weighted components:

- Labeled instances are treated as positives with weight $1/e(x|\mathbb{K}_x)$, reflecting our belief that this observed positive represents only a fraction of a larger group of similar true positives.
- Unlabeled instances are treated as negatives with weight 1, under the naive assumption of being non-positive examples.
- Labeled instance are also included as a pseudo-negative with negative weight $1 - 1/e(x|\mathbb{K}_x) \leq 0$, to counterbalance the assumption that all unlabeled instances are negative despite the expectation that some of them are actually positive.

This dual weighting mechanism encourages the model to identify positive instances while reducing the penalty for assigning high confidence to unlabeled instances that resemble known positives. Here, "similarity" extends beyond feature-based similarity to include graph-theoretic properties, such as structural roles, local neighborhoods, and global positions within the observed entity-graph.

We observe that Supplementary Equation SM53 extends the approach proposed by Wardeh et al. [16] for predicting missing mammal-virus associations by not only addressing class imbalance but also incorporating epistemic uncertainty

about the true class of unlabeled instances. This is achieved through the introduction of negative weights applied to labeled instances. Moreover, our formulation defines the weights in a principled way based on the inverse of their dynamic propensity scores, instead of relying on heuristic or fixed weights schemes.

Non-negative empiric DPU risk

The minimization of Supplementary Equation SM53 can be used to train our model g_θ in an unbiased way using only labeled and unlabeled instances in \mathbb{K} . However, as noticed by Kiryo et al. [41], flexible models trained under a PU setting are prone to overfitting when minimizing the empirical PU risk. This often leads to pathological behavior, such as excessively negative risk values and poor generalization performance. The same issue arises in the DPU setting. Kiryo et al. observed that for any classifier g , the expected loss over negative instances is non-negative:

$$\mathbb{E}_{\mathcal{N}}[\ell(g(x), 0)] = \pi_s \mathbb{E}_{\mathcal{S}} \left[\left(1 - \frac{1}{e(x)}\right) \ell(g(x), 0) \right] + (1 - \pi_s) \mathbb{E}_{\mathcal{U}}[\ell(g(x), 0)] \geq 0, \quad (\text{SM54})$$

To address this, they proposed a regularization method that ensures a non-negative empirical risk during training. We adopt a similar strategy for our DPU risk formulation, resulting in the non-negative DPU empirical risk $\hat{R}_{\ell, \mathbb{K}, \mathcal{M}}^{\text{NNDPU}}$, defined as:

$$\hat{R}_{\ell, \mathbb{K}, \mathcal{M}}^{\text{NNDPU}}[g] := \hat{R}_{\ell, \mathbb{K}, \mathcal{M}}^{\mathcal{S}, +}[g] + \max \left\{ -\beta, \hat{R}_{\ell, \mathbb{K}, \mathcal{M}}^{\mathcal{S}, -}[g] + \hat{R}_{\ell, \mathbb{K}, \mathcal{M}}^{\mathcal{U}, -}[g] \right\}, \quad (\text{SM55})$$

with $\beta \geq 0$ is a hyperparameter to be tuned during training. The individual components of the risk are defined as:

$$\begin{aligned} \hat{R}_{\ell, \mathbb{K}, \mathcal{M}}^{\mathcal{S}, +}[g] &= \frac{1}{|\Omega| \times |\mathcal{M}|} \sum_{x \in \mathcal{S}(\mathbb{K})} \sum_{\Lambda \in \mathcal{M}(x)} \frac{1}{e(x; \mathbb{K}_{\{\Lambda\}})} \ell(g(x; \mathbb{K}_{\Lambda}; \mathbf{X}_g), 1), \\ \hat{R}_{\ell, \mathbb{K}, \mathcal{M}}^{\mathcal{S}, -}[g] &= \frac{1}{|\Omega| \times |\mathcal{M}|} \sum_{x \in \mathcal{S}(\mathbb{K})} \sum_{\Lambda \in \mathcal{M}(x)} \left(1 - \frac{1}{e(x; \mathbb{K}_{\{\Lambda\}})}\right) \ell(g(x; \mathbb{K}_{\Lambda}; \mathbf{X}_g), 0), \\ \hat{R}_{\ell, \mathbb{K}, \mathcal{M}}^{\mathcal{U}, -}[g] &= \frac{1}{|\Omega| \times |\mathcal{M}|} \sum_{x \in \mathcal{U}(\mathbb{K})} \sum_{\Lambda \in \mathcal{M}(x)} \ell(g(x; \mathbb{K}_{\Lambda}; \mathbf{X}_g), 0). \end{aligned}$$

This formulation ensures robust training by preventing the empirical risk from becoming arbitrarily negative, even when using highly expressive models.

3.2 Graph-dataset creation

In this section, we describe a method for constructing a dataset of graphs suitable for training graph-based models in a DPU setting.

We focus on the realistic scenario in which only the current configuration \mathbb{K} of the database is available. This reflects our situation with the VIRION database, where temporal information about when specific mammal-virus associations were discovered is either unavailable or inconsistently recorded. Such limitations are common in many real-world applications, where historical snapshots of the database are not maintained.

Also, for simplicity, we assume that our \mathbb{K} does not include a matrix feature \mathbf{X}_k . Specifically, in the current stage of our work, we are not considering any matrix feature \mathbf{X}_k for observed mammal-virus associations, although incorporating such features may be beneficial in future studies.

Graph-datasets for training and evaluation

A widely used approach for training a GNN model to predict new links in an incomplete graph involves using a set of K subgraphs derived from the original graph $G(\mathbb{K}) = (\mathcal{V}, E = \mathcal{S}(\mathbb{K}))$ by masking a part of the original edges.

To this hand, we define a graph-dataset based on \mathbb{K} as a collection of $K \in \mathbb{N}$ subgraphs $G^{(k)}$ of $G(\mathbb{K})$, each associated with a subset of labeled associations $\mathcal{E}_{\mathcal{S}}^{(k)} \subset \mathcal{S}(\mathbb{K})$ and a subset of unlabeled associations $\mathcal{E}_{\mathcal{U}}^{(k)} \subset \mathcal{U}(\mathbb{K})$, which are used as target edges during training or evaluation:

$$\mathcal{D} = \left\{ \left(G^{(k)}, \mathcal{E}_{\mathcal{S}}^{(k)}, \mathcal{E}_{\mathcal{U}}^{(k)} \right) \right\}_{k=1}^K. \quad (\text{SM56})$$

Each subgraph $G^{(k)} = (\mathcal{V}, E^{(k)} \subset \mathcal{S}(\mathbb{K}))$ is constructed by removing the labeled target edges $\mathcal{E}_{\mathcal{S}}^{(k)}$ from the original graph, i.e.,

$$E^{(k)} = \mathcal{S}(\mathbb{K}) \setminus \mathcal{E}_{\mathcal{S}}^{(k)}, \quad \forall k \in \{1, \dots, K\}. \quad (\text{SM57})$$

Algorithm for partitioning associations

To construct the graph-datasets, we partition the labeled and unlabeled associations in $\mathcal{S}(\mathbb{K})$ and $\mathcal{U}(\mathbb{K})$ using Algorithm 2, resulting in the subsets $\mathcal{E}_{\mathcal{S}}^{(k)}$ and $\mathcal{E}_{\mathcal{U}}^{(k)}$ for each $k \in \{1, \dots, K\}$. The goal of the algorithm is to evenly partition

a given set of mammal-virus associations $\mathcal{E} \subseteq M \times V$ into K balanced non-overlapping folds $\mathcal{E}^{(k)}$. The pseudocode of Algorithm 2 is found in the Supplementary Note section.

The algorithm ensures that associations are balanced across folds, with each fold containing a similar number of instances with comparable characteristics. These characteristics include the viral family for nodes $v \in V$ and the degree, mammalian order, and mammalian family for nodes $m \in M$.

This approach aims to maintain homogeneity in the statistical properties of the associations across folds, minimizing the risk of any fold having disproportionately more or fewer associations from a specific category. Also, the algorithm can be generalized to other datasets by replacing the constraints related to viral and mammal node classes with other relevant node classes specific to the problem at hand.

3.3 Our training approach

We conducted multiple model trainings, each tailored to serve a specific objective. First, we performed a 10-fold cross-validation using only labels from the VIRION database to evaluate model robustness and tune the hyperparameter used for all subsequent training procedures. Second, we trained a single model using all VIRION labels to assess prediction performance on new labels obtained from the GenBank database. Third, we trained a single model combining VIRION labels with synthetic association labels to evaluate the model’s ability to recover known ground truth for unlabeled synthetic associations. Finally, to produce the most up-to-date predictions for all currently unlabeled associations and to quantify predictive uncertainty, we trained an ensemble of models using combined labels from both the VIRION and GenBank databases.

Propensity score estimation

Since we are not provided with the propensity scores of mammal-virus associations, we estimate them using the Bayesian DPU model (Subsection 2.4). These informed estimates served two purpose: first, to derive the empirical DPU risk (Supplementary Equation SM59) used to train the classifier model (Subsection 2.5); second, they are integrated in the final prediction step (Supplementary Equation SM12) to compute the final probabilities for unlabeled associations.

For 10-fold cross-validation, predicting new labels, and evaluating predictions on synthetic data, we computed the estimated propensity scores $\hat{e}_{m,v}$ as the expected values of the propensity score $e_{\Phi}(m, v; \mathbb{K}_{\{(m,v)\}})$ under the posterior distribution of Φ_v and Π_v :

$$\hat{e}_{m,v} = \int d\Phi_v d\Pi_v \Pr(\Phi_v, \Pi_v | \mathbb{K}) e_{\Phi}(m, v; \mathbb{K}_{\{(m,v)\}}). \quad (\text{SM58})$$

In these cases, a single set of propensity scores is generated and a single classifier is trained.

Instead, for the final set of predictions—where we combine labels from both VIRION and GenBank to produce the most up-to-date estimates—we adopted an ensemble-based strategy to incorporate the full posterior uncertainty of the Bayesian model. Specifically, we generated $N = 100$ ensemble members, where each classifier $g^{(i)}$ was trained using a distinct graph-dataset and set of propensity scores (Supplementary Notes 4). These propensity scores were computed using independent samples of the Bayesian DPU model parameters Φ_v drawn from their posterior distribution. Predictions were then obtained by combining the outputs of classifiers $g^{(i)}$ with their corresponding set of propensity scores via Supplementary Equation SM12.

Non-negative empiric DPU risk for mammal-virus associations

Given a graph-dataset $\mathcal{D} = \left\{ \left(G^{(k)}, \mathcal{E}_S^{(k)}, \mathcal{E}_U^{(k)} \right) \right\}_{k=1}^K$, the classifier is trained by minimizing the following empiric non-negative empiric DPU risk:

$$\hat{R}_{\ell}^{\text{DPU}}[g; \mathcal{D}] := \frac{1}{|V|} \sum_{v \in V} \left[\hat{R}_{\ell,v}^{S,+}[g; \mathcal{D}] + \max \left\{ 0, \hat{R}_{\ell,v}^{S,-}[g; \mathcal{D}] + \hat{R}_{\ell,v}^{U,-}[g; \mathcal{D}] \right\} \right]. \quad (\text{SM59})$$

Its components are:

$$\begin{aligned} \hat{R}_{\ell,v}^{S,+}[g; \mathcal{D}] &= \frac{\pi_{s,v}}{\sum_{k=1}^K |\mathcal{E}_{S,v}^{(k)}|} \sum_{k=1}^K \sum_{x \in \mathcal{E}_{S,v}^{(k)}} \frac{1}{\hat{e}_x} \ell \left(g(x, G^{(k)}), 1 \right), \\ \hat{R}_{\ell,v}^{S,-}[g; \mathcal{D}] &= \frac{\pi_{s,v}}{\sum_{k=1}^K |\mathcal{E}_{S,v}^{(k)}|} \sum_{k=1}^K \sum_{x \in \mathcal{E}_{S,v}^{(k)}} \left(1 - \frac{1}{\hat{e}_x} \right) \ell \left(g(x, G^{(k)}), 0 \right), \\ \hat{R}_{\ell,v}^{U,-}[g; \mathcal{D}] &= \frac{1 - \pi_{s,v}}{\sum_{k=1}^{K_{\text{train}}} |\mathcal{E}_{U,v}|} \sum_{k=1}^K \sum_{x \in \mathcal{E}_{U,v}^{(k)}} \ell \left(g(x, G^{(k)}), 0 \right), \end{aligned}$$

with $\mathcal{E}_{S,v}^{(k)} = \mathcal{E}_S^{(k)} \cap \{\cup_{m \in M}(m, v)\}$ and $\mathcal{E}_{U,v}^{(k)} = \mathcal{E}_U^{(k)} \cap \{\cup_{m \in M}(m, v)\}$, while $\pi_{s,v} = |\mathcal{S}(\mathbb{K}) \cap \{\cup_{m \in M}(m, v)\}| / |\Omega|$. Also, the loss function $\ell : [0, 1] \times [0, 1] \rightarrow \mathbb{R}^+$ used is the binary cross entropy loss function, defined as:

$$\ell(\hat{y}, y) = -y \log(\hat{y}) - (1 - y) \log(1 - \hat{y}), \quad (\text{SM60})$$

which is typically adopted for training binary classifiers.

Importantly, non-negativity is enforced individually for each viral family $v \in V$, which provides stronger regularization than applying it after aggregation. Given a labeled association (m, v) , this strategy prevents the classifier to overestimate associations (m, v') for $v' \neq v$ merely due to the presence of a labeled association with v . Instead, it encourages the recall of associations (m', v) with $m' \neq m$. In this way, the presence of a labeled association (m, v) promotes higher predictions across mammal species for the same viral family, rather than across viral families for the same mammal species.

10-fold cross-validation

To tune hyperparameters and evaluate model robustness, we employ a 10-fold cross-validation scheme with rotating validation and test sets. First, using Algorithm 2, labeled and unlabeled associations are evenly partitioned into 10 non-overlapping folds:

$$\mathcal{E}_S^{(i)} \subset \mathcal{S}(\mathbb{K}) \quad \text{and} \quad \mathcal{E}_U^{(i)} \subset \mathcal{U}(\mathbb{K}), \quad \text{with } i \in \{1, \dots, 10\}. \quad (\text{SM61})$$

For each iteration $i \in \{1, \dots, 10\}$, one fold is used for testing, another for validation (i.e., model selection), and the remaining eight folds are aggregated for training. This ensures that across the 10 iterations, each fold serves exactly once as test set, once as validation set, and eight times as part of the training set. As a result, each iteration uses 80% of the associations for training, 10% for model selection, and 10% for final evaluation on a held-out set that remains unseen during both training and model selection.

For each iteration $i \in \{1, \dots, 10\}$, associations in each fold are further partitioned using Algorithm 2. Associations in test and validation sets are divided into $K_{\text{test}} = K_{\text{val}} = 50$ balanced subsets, while associations in training set are divided into $K_{\text{train}} = (10 - 2) \times 50 = 400$ subsets, i.e.,

$$\begin{aligned} \mathcal{E}_{\text{test}, S}^{(i,k)} &\subset \mathcal{E}_{\text{test}, S}^{(i)} & \text{and} & & \mathcal{E}_{\text{test}, U}^{(i,k)} &\subset \mathcal{E}_{\text{test}, U}^{(i)}, & \text{for } k \in \{1, \dots, K_{\text{test}}\}, \\ \mathcal{E}_{\text{val}, S}^{(i,k)} &\subset \mathcal{E}_{\text{val}, S}^{(i)} & \text{and} & & \mathcal{E}_{\text{val}, U}^{(i,k)} &\subset \mathcal{E}_{\text{val}, U}^{(i)}, & \text{for } k \in \{1, \dots, K_{\text{val}}\}, \\ \mathcal{E}_{\text{train}, S}^{(i,k)} &\subset \mathcal{E}_{\text{train}, S}^{(i)} & \text{and} & & \mathcal{E}_{\text{train}, U}^{(i,k)} &\subset \mathcal{E}_{\text{train}, U}^{(i)}, & \text{for } k \in \{1, \dots, K_{\text{train}}\}. \end{aligned} \quad (\text{SM62})$$

These partitions are used for creating graph-datasets. For example, edge set of subgraph $G_{\text{test}}^{(i,k)}$ is given by:

$$E_{\text{test}}^{(i,k)} = \mathcal{S}(\mathbb{K}) \setminus \mathcal{E}_{\text{test}, S}^{(i,k)}, \quad \forall k \in \{1, \dots, K_{\text{test}}\}. \quad (\text{SM63})$$

The same procedure is applied to generate validation and training subgraphs, using their respective partitions. This results in subgraphs with approximately 10 missing edges each with respect to the original graph $G(\mathbb{K})$.

Repeating this subgraph creation $N = 5$ times per fold results in $N \times K_{\text{test}} = N \times K_{\text{val}} = 250$ subgraphs for the test and validation graph-datasets, and $N \times K_{\text{train}} = 2000$ subgraphs for the training graph-dataset.

The graph-datasets for each iteration i are:

$$\begin{aligned} \mathcal{D}_{\text{test}}^{(i)} &= \left\{ (G_{\text{test}}^{(i,k)}, \mathcal{E}_{\text{test}, S}^{(i,k)}, \mathcal{E}_{\text{test}, U}^{(i,k)}) \right\}_{k=1}^{N \times K_{\text{test}}}, & \forall i \in \{1, \dots, 10\}, \\ \mathcal{D}_{\text{val}}^{(i)} &= \left\{ (G_{\text{val}}^{(i,k)}, \mathcal{E}_{\text{val}, S}^{(i,k)}, \mathcal{E}_{\text{val}, U}^{(i,k)}) \right\}_{k=1}^{N \times K_{\text{val}}}, & \forall i \in \{1, \dots, 10\}, \\ \mathcal{D}_{\text{train}}^{(i)} &= \left\{ (G_{\text{train}}^{(i,k)}, \mathcal{E}_{\text{train}, S}^{(i,k)}, \mathcal{E}_{\text{train}, U}^{(i,k)}) \right\}_{k=1}^{N \times K_{\text{train}}}, & \forall i \in \{1, \dots, 10\}. \end{aligned} \quad (\text{SM64})$$

Since in this way each mammal-virus association appears in the graph-dataset as a target edge $N = 5$ times, it is predicted $N = 5$ times, each time using a different subgraph. Thus, in this case, we defined the model's final prediction as the average of the predictions obtained from these subgraphs.

Training using all associations

For predicting new labels and evaluating predictions on synthetic data, we trained a single classifier using all labeled and unlabeled associations in the database considered \mathbb{K} —either the VIRION database (for new label prediction) or VIRION augmented with synthetic associations (for synthetic evaluation).

Labeled and unlabeled associations are partitioned into $K = 500$ non-overlapping subsets using Algorithm 2. For each partition $k \in \{1, \dots, K\}$, a subgraph $G^{(k)} = (\mathcal{V}, E^{(k)} \subset \mathcal{S}(\mathbb{K}))$ is created by retaining all labeled associations in $G(\mathbb{K})$ that are not in $\mathcal{E}_S^{(k)}$, that is:

$$E^{(k)} = \mathcal{S}(\mathbb{K}) \setminus \mathcal{E}_S^{(k)}, \quad \forall k \in \{1, \dots, K\}. \quad (\text{SM65})$$

To enhance model robustness, the partitioning process is repeated $N = 5$ times generating a total of $N \times K = 2500$ subgraphs. This results in the graph-dataset:

$$\mathcal{D} = \left\{ \left(G^{(k)}, \mathcal{E}_S^{(k)}, \mathcal{E}_U^{(k)} \right) \right\}_{k=1}^{N \times K}. \quad (\text{SM66})$$

Given that $\mathcal{S}(\mathbb{K})$ contains $\sim 5,000$ associations, each subgraph $G^{(k)}$ will, on average, have ~ 10 missing edges with respect to $G(\mathbb{K})$. Also in this case, since each mammal-virus association is predicted using $N = 5$ different subgraphs, we define the model’s final prediction as the average of the predictions obtained from these subgraphs.

To generate up-to-date predictions, each ensemble classifier is trained and evaluated on a distinct graph-dataset constructed using the procedure above. This diversity enables us to estimate prediction uncertainty with respect to the dataset variability.

Hyperparameters and model regularization

We optimized the model using the AdamW optimizer [42, 43] with an initial learning rate of 10^{-3} and weight decay of 10^{-3} . Training is conducted over 2000 epochs with a batch size of 50 subgraphs. Mixed precision training was employed to improve computational efficiency [44].

Regularization beyond weight decay was applied via two custom penalties: the first penalized the model when embeddings \mathbf{z}_m and \mathbf{z}_v (Supplementary Equation SM51) of the same node differ across subgraphs in a batch, encouraging consistency via feature-wise cosine similarity; the second penalized variance in predictions for the same unlabeled associations across a batch’s subgraphs. The regularization strength λ for both penalties started at 0 and was gradually increased to 10 between epochs 250 and 800.

Gradient clipping with a maximum norm of 0.3 was applied to stabilize training. Additionally, inverse propensity score weights in the loss function (Equation SM59) were clipped at 500, corresponding to a minimum propensity score of 0.002, to mitigate instability caused by excessively large weights. On average, only about 10 instances exceeded this threshold. This clipping helps reduce the variance of the DPU loss function, leading to more stable outcomes, albeit at the cost of introducing some bias [45]. Consequently, on average, the model tends to recall fewer unobserved associations as true. However, given the relatively high clipping threshold, the resulting bias is expected to be limited, with more conservative predictions primarily affecting only the most understudied groups of mammal–virus interactions.

Training for 2000 epochs ensured that the loss function and key metrics (Naive-Recall, Naive-NegRecall, Naive-AUC, Naive-PRAUC, and SAR-PUF₁; see Section 2 of Supplementary Notes) plateau. For 10-fold cross-validation, the model with the best SAR-PUF₁ on $\mathcal{D}_{\text{val}}^{(i)}$ is selected and used to evaluate performance on $\mathcal{D}_{\text{test}}^{(i)}$. When training on all labeled data, the model with the highest SAR-PUF₁ on the training set is selected.

4 Uncertainty calculation, aggregated predictions and limitations

The ensemble-based strategy (reported schematically in Supplementary Figure SM6) for generating final predictions allows us to incorporate and propagate the Bayesian uncertainty in the estimated propensity scores into the final predictions. The ensemble consisted of $N = 100$ independently trained models.

4.1 Uncertainty calculation

The ensemble methods allows us to effectively treat the probability that a certain potential association $x \in M \times V$ exists as a random variable sampled from the set of ensemble outputs:

$$p_x \sim \text{Uniform} \left(\left\{ p_x^{(1,1)}, \dots, p_x^{(1,K)}, \dots, p_x^{(i,k)}, \dots, p_x^{(N,1)}, \dots, p_x^{(N,K)} \right\} \right), \quad (\text{SM67})$$

where $p_x^{(i,k)}$ represents the probability assigned to x by the ensemble member $i \in \{1, \dots, N = 100\}$ on graph dataset $\mathcal{D}^{(i,k)}$, with $k \in \{1, \dots, K = 5\}$.

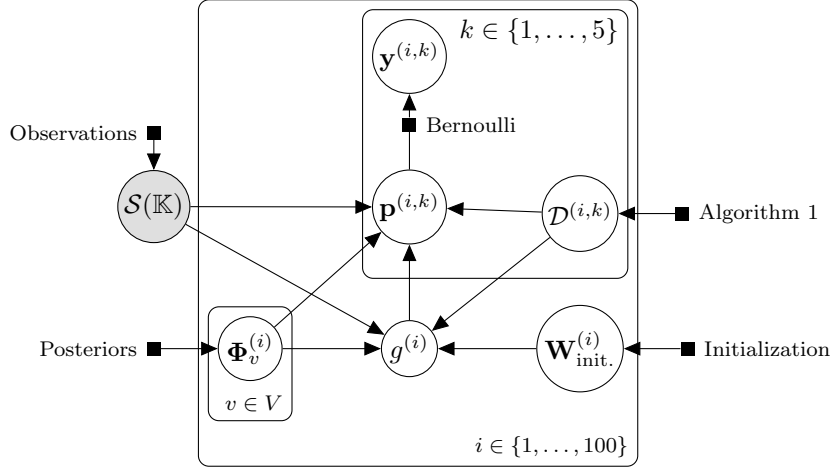
Each ensemble member i consists of a trained GNN classifier $g^{(i)}$, and a set of propensity scores parameters $\{\Phi_v^{(i)}\}_{v \in V}$, sampled from the posterior distribution of the Bayesian DPU model.

The actual class y_x is then modeled as a Bernoulli random variable conditioned on the sampled probability p_x :

$$y_x \sim \text{Bernoulli}(p_x). \quad (\text{SM68})$$

This two-stage sampling process captures multiple sources of uncertainties:

- *model uncertainty*, arising from differences in propensity score samples across ensemble members;
- *aleatoric uncertainty*, due to the inherent randomness in the label assignment given a predicted probability p_x ;
- *dataset-induced uncertainty*, reflecting how the choice of graph dataset $\mathcal{D}^{(i,k)}$ used for evaluation can influence predictions by exposing the model to different observed contexts.



Supplementary Figure SM6 | Graphical model of the ensemble-based strategy used to propagate uncertainty from Bayesian propensity score estimation to final predictions. For each ensemble member $i \in \{1, \dots, 100\}$, we sample parameters $\Phi_v^{(i)}$ from the posterior distribution of the hierarchical Bayesian model and compute the corresponding propensity scores $\{e_i(m, v)\}_{(m,v) \in M \times V}$. These scores, along with the association labels $\mathcal{S}(\mathbb{K})$ and a randomly generated graph-based dataset $\mathcal{D}^{(i,k)}$ (constructed using Algorithm 2), are used to train a classifier $g^{(i)}$ with randomly initialized weights $\mathbf{W}_{\text{init}}^{(i)}$. The index $k \in \{1, \dots, 5\}$ here is to make explicit that each classifier $g^{(i)}$ is associated with 5 distinct graph-dataset, created by repeating Algorithm 2 5 times for ensemble member. By combining the labels of the instances, their sampled propensity scores, and the classifier outputs on the graph-dataset $\mathcal{D}^{(i,k)}$, we compute class probabilities $\mathbf{p}^{(i,k)} = \{p_{i,k}(m, v)\}_{(m,v) \in M \times V}$ via Equation (SM12), for each i and k . Final predictions $\mathbf{y}^{(i,k)} = \{y_{i,k}(m, v)\}_{(m,v) \in M \times V}$ are then sampled from Bernoulli distributions parameterized by $\mathbf{p}^{(i,k)}$. For clarity, we omitted the explicit dependency of both the posteriors over Φ_v and the datasets $\mathcal{D}^{(i,k)}$ on the labels $\mathcal{S}(\mathbb{K})$.

4.2 Aggregated predictions

Given an ensemble member i and a graph dataset index k , the expected value of the aggregated quantity is given by:

$$y_{\text{aggr}}^{(i,k)} = \sum_{x \in A} y_x^{(i,k)}, \quad (\text{SM69})$$

where $A \subset M \times V$ represents a subset of potential associations grouped by a common feature (e.g., all associations related to a particular mammal species, order, or viral family). The sum of the classes y_x over $x \in A$ corresponds to the total number of predicted associations within that group. Its expected value can be expressed as:

$$\mathbb{E}[y_{\text{aggr}}^{(i,k)}] = \mathbb{E}\left[\sum_{x \in A} y_x^{(i,k)}\right] = \sum_{x \in A} \mathbb{E}[y_x^{(i,k)}] = \sum_{x \in A} p_x^{(i,k)}, \quad (\text{SM70})$$

where the equality follows from the linearity of expectation.

To summarize the aggregated prediction across the entire ensemble, we take the median of these expected aggregated values $\mathbb{E}[y_{\text{aggr}}^{(i,k)}]$ over all ensemble members i and graph dataset indexes k , along with the 5th and 95th percentiles to reflect ensemble variability in the expected number of predicted associations, that is:

$$y_{\text{aggr}} = \text{median}_{(i,k)} \mathbb{E}[y_{\text{aggr}}^{(i,k)}], \quad (\text{SM71})$$

$$\text{CI}(y_{\text{aggr}}) = [\text{quantile}(0.05)_{(i,k)} \mathbb{E}[y_{\text{aggr}}^{(i,k)}]; \text{quantile}(0.95)_{(i,k)} \mathbb{E}[y_{\text{aggr}}^{(i,k)}]]. \quad (\text{SM72})$$

Some notes on the uncertainties of aggregated predictions

It is important to clarify that the confidence intervals reported for aggregated predictions, as defined in Supplementary Equation SM72, reflect only the variability in expected values across ensemble members. These intervals do not represent full confidence intervals over the actual aggregated predictions.

Obtaining true confidence intervals for aggregated predictions would require computing joint probabilities across multiple associations. However, this is infeasible, as it involves evaluating complex conditional distributions that do not factorize into products of marginal probabilities. In fact, while the expectation of a sum equals the sum of expectations, this does not hold for variances or confidence intervals when associations are statistically dependent. This dependency is particularly relevant in the DPU setting, where predictions for different associations are not

conditionally independent given the model parameters. The DPU model explicitly relies on the idea that what is known about some associations influences the likelihood of others. For instance, if a virus is predicted to infect one host, it becomes more likely to infect related hosts—and less likely if not. These positive correlations mean that uncertainties across associations co-vary rather than cancel out, leading to systematic underestimation of uncertainty when independence is assumed.

Therefore, the reported confidence intervals should be interpreted as reflecting variation in expected values across the ensemble—driven by differences in propensity scores, model weight initializations, and graph samples—not as full measures of predictive uncertainty. In particular, they do not account for aleatoric uncertainty or correlations between individual predictions within the aggregated set. However, while coming with limitations, these intervals still offer meaningful insight into how the expected number of predicted associations varies across different model realizations.

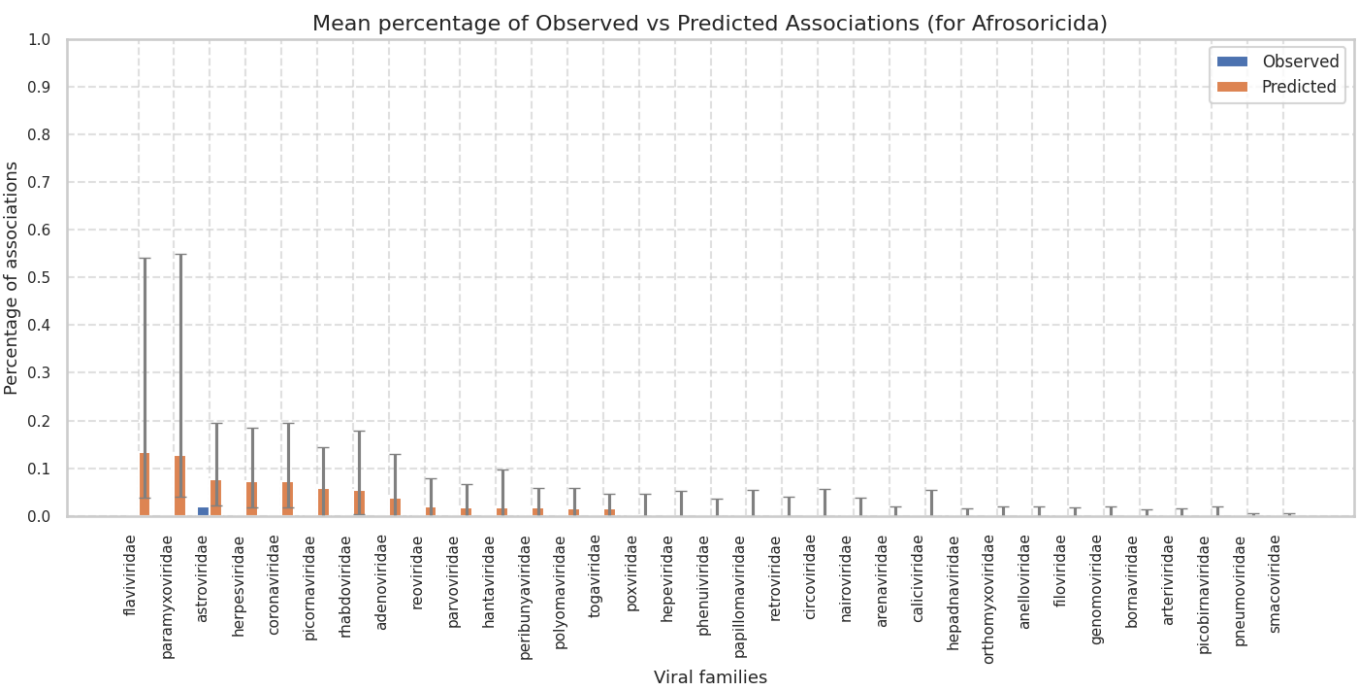
4.3 Limitations of uncertainty estimates

It is important to emphasize that the uncertainties computed for individual associations do not capture uncertainty under potential new data acquisition. Specifically, the DPU loss (SM59) encourages the model to assign higher probabilities to unlabeled associations that resemble known positives, while simultaneously discouraging positive predictions for associations that are dissimilar or isolated in the feature space. As a result, highly isolated associations—those that are distant from any labeled examples—tend to receive consistently low predicted probabilities from all classifiers in the ensemble. This behavior persists even when such associations are coupled with low estimated propensity scores. Therefore, the predicted probabilities and associated uncertainties for such under-sampled or poorly studied groups of associations are likely to be underestimated, potentially overlooking genuine but undocumented associations. This limitation should be considered when interpreting predictions in severely underexplored regions of the mammal–virus association space. A more comprehensive treatment of predictive uncertainty—especially accounting for hypothetical future data acquisition—is beyond the scope of the present work.

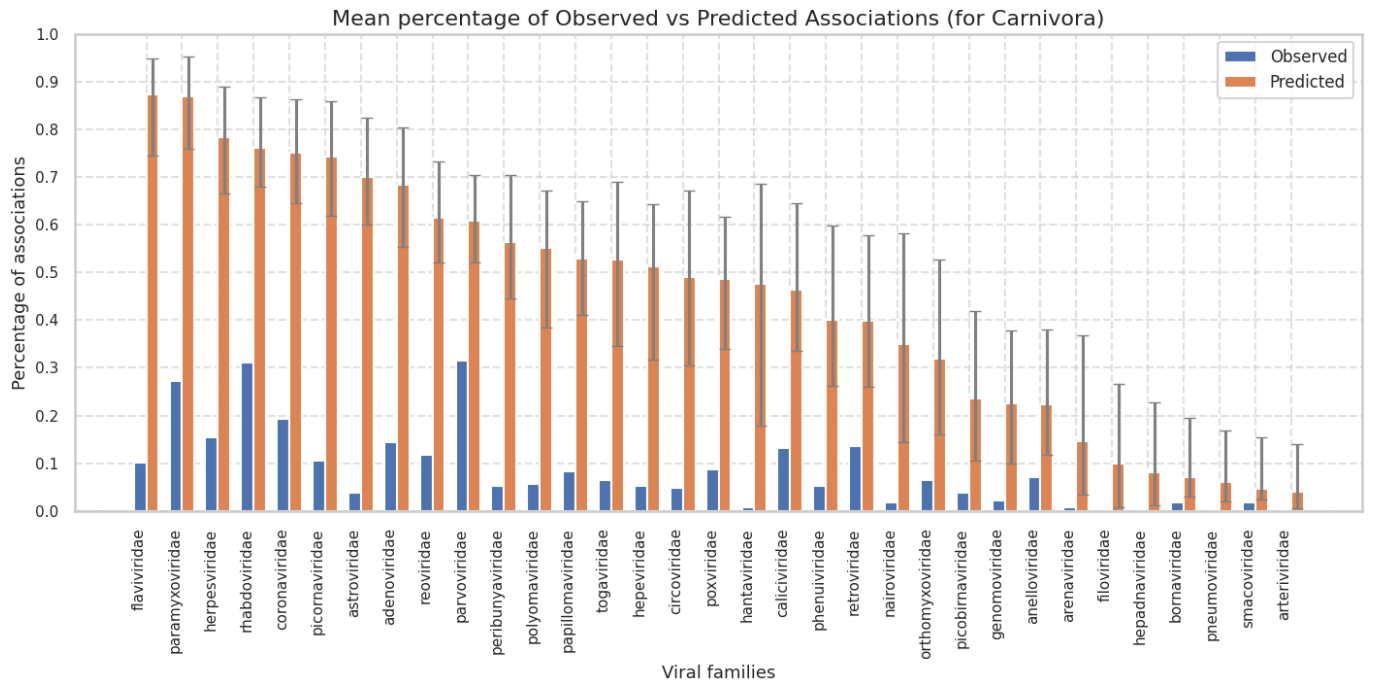
Supplementary Results

1 Predictions for each mammalian orders

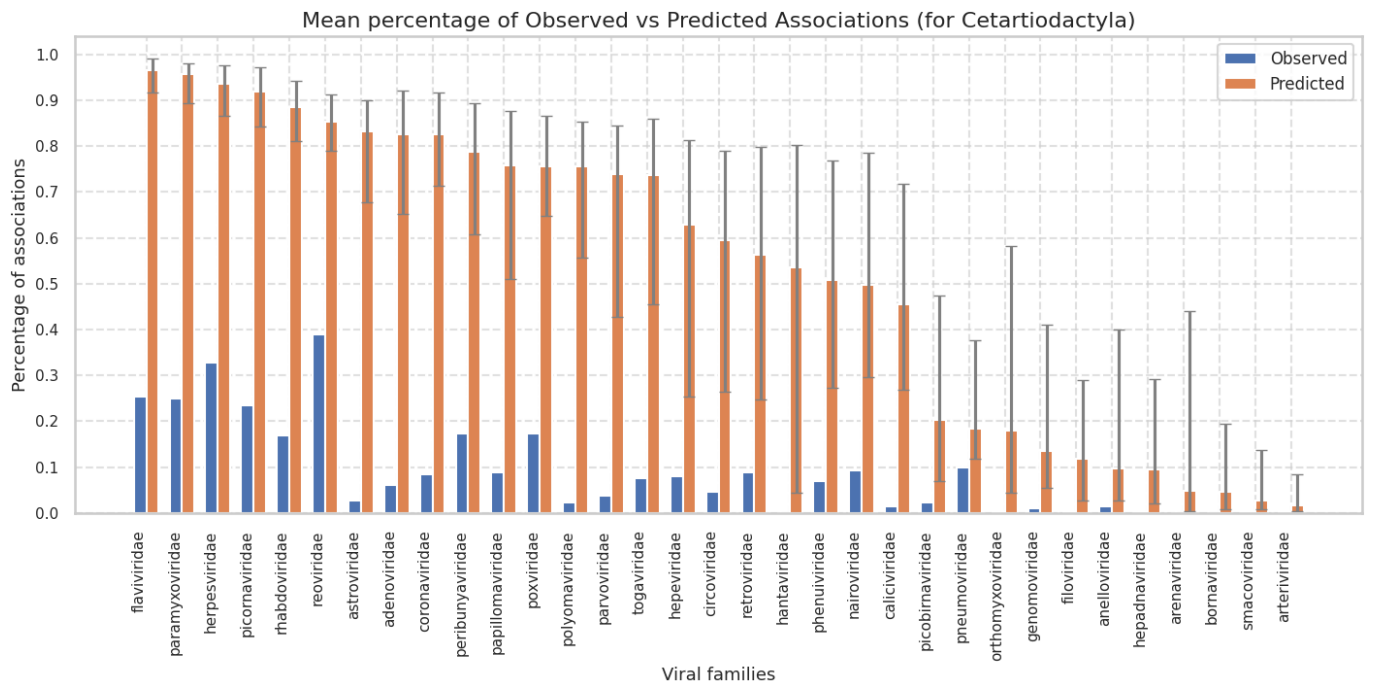
Here, for each mammalian order, we show the percentage of observed vs predicted associations across viral families (Supplementary Figures SR1–SR26).



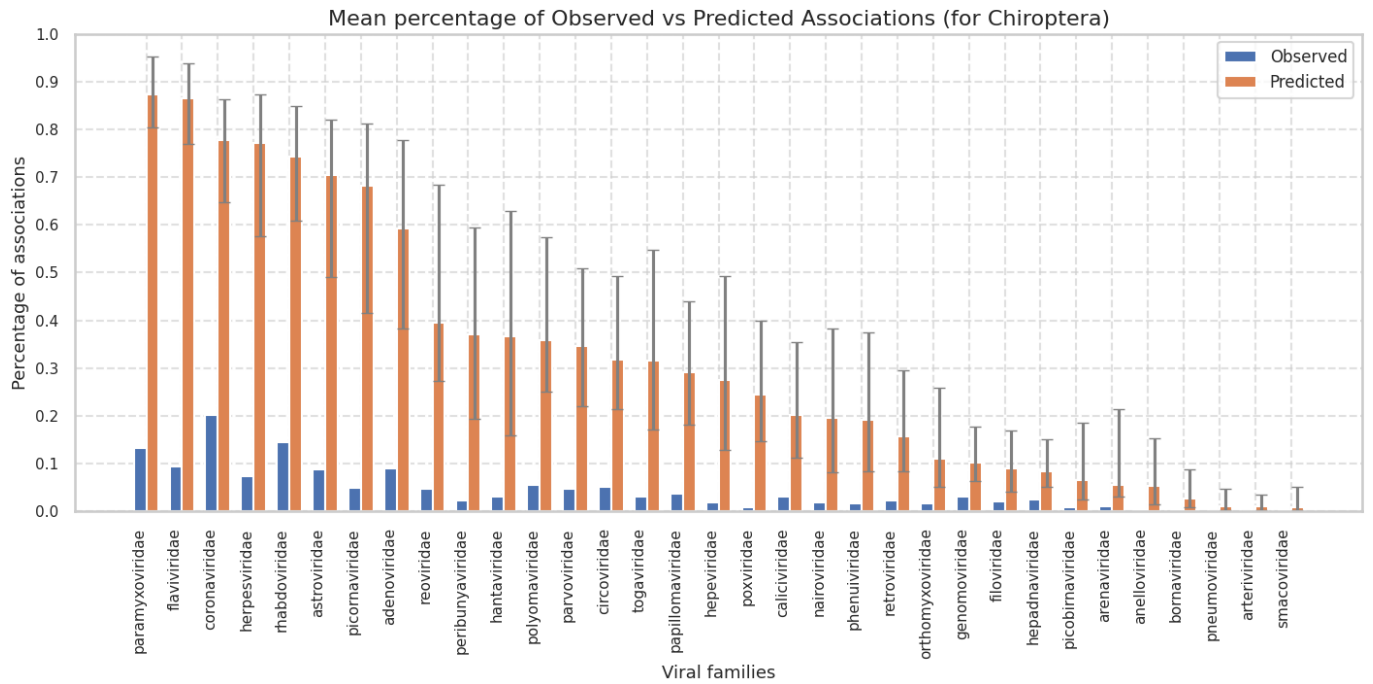
Supplementary Figure SR1 | Predicted associations by viral family for Afrosoricida. Percentage of predicted associations across viral families. Orange bars represent the medians of expected percentages across ensemble members, with uncertainty bars indicating the 5th to 95th percentile range. Blue bars indicate the observed percentages.



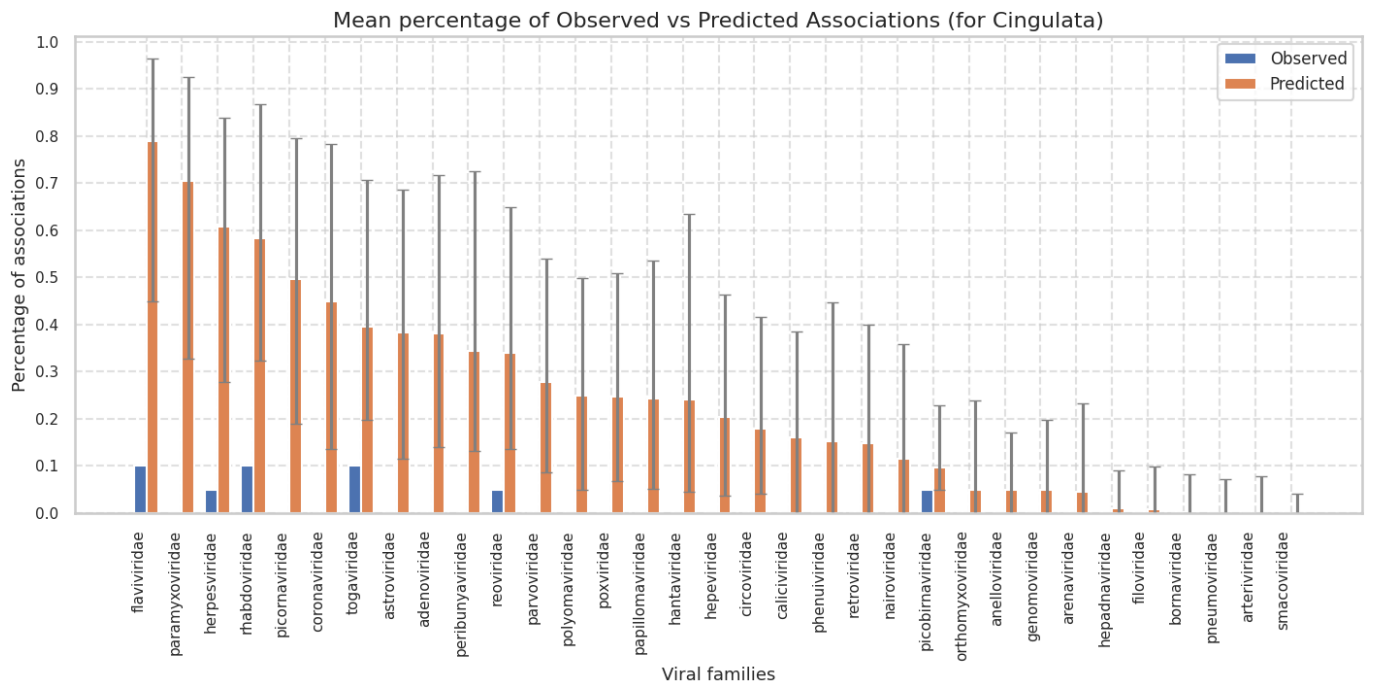
Supplementary Figure SR2 | Predicted associations by viral family for Carnivora. Percentage of predicted associations across viral families. Orange bars represent the medians of expected percentages across ensemble members, with uncertainty bars indicating the 5th to 95th percentile range. Blue bars indicate the observed percentages.



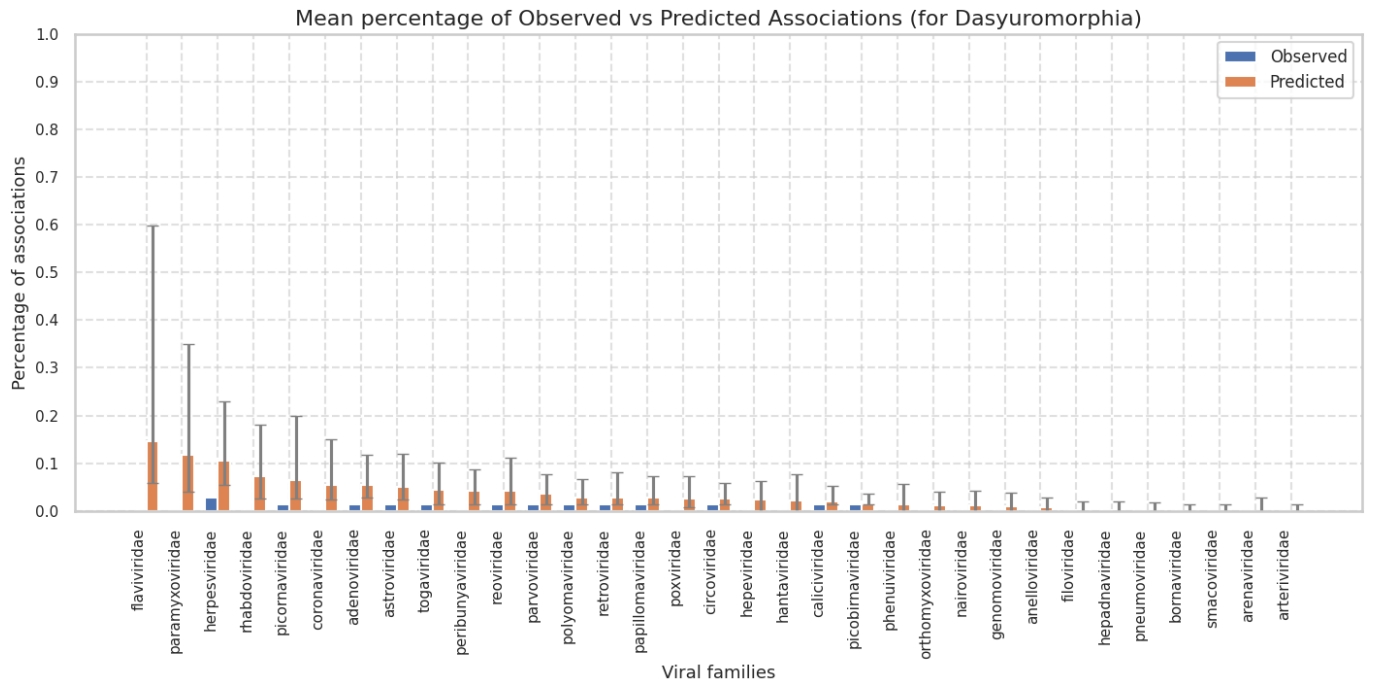
Supplementary Figure SR3 | Predicted associations by viral family for Cetartiodactyla. Percentage of predicted associations across viral families. Orange bars represent the medians of expected percentages across ensemble members, with uncertainty bars indicating the 5th to 95th percentile range. Blue bars indicate the observed percentages.



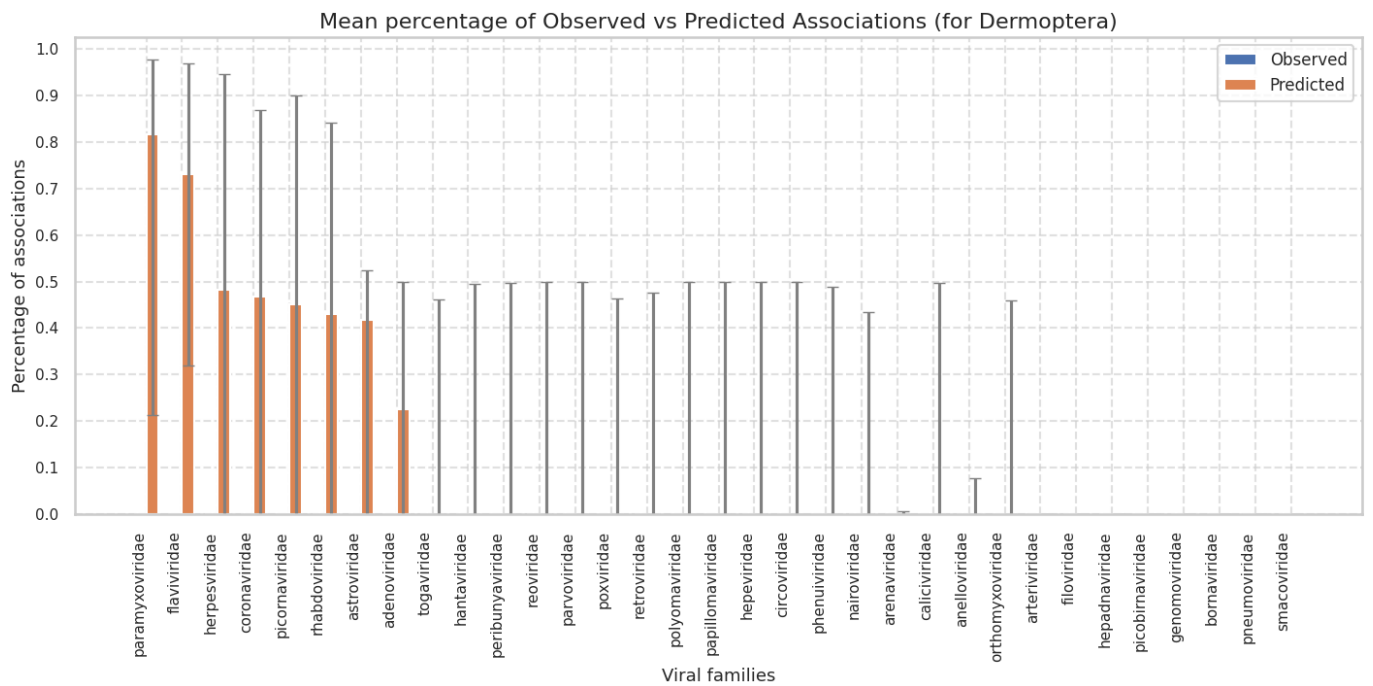
Supplementary Figure SR4 | Predicted associations by viral family for Chiroptera. Percentage of predicted associations across viral families. Orange bars represent the medians of expected percentages across ensemble members, with uncertainty bars indicating the 5th to 95th percentile range. Blue bars indicate the observed percentages.



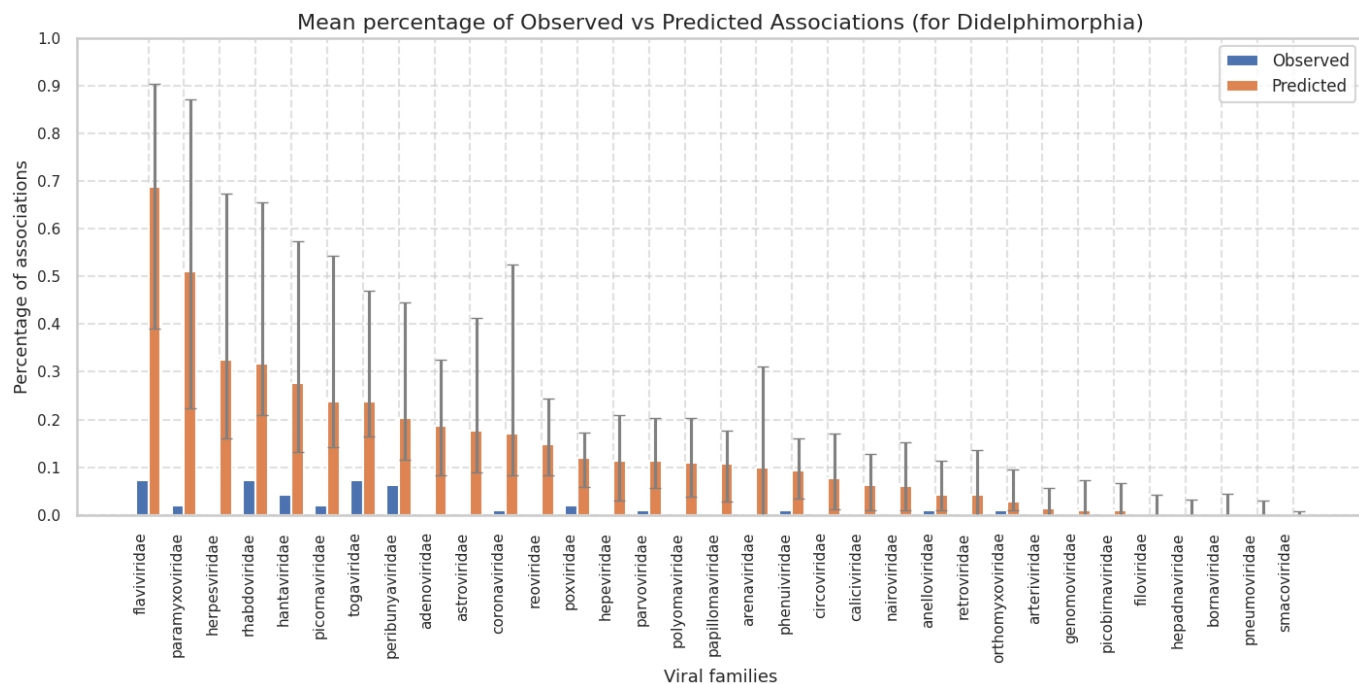
Supplementary Figure SR5 | Predicted associations by viral family for Cingulata. Percentage of predicted associations across viral families. Orange bars represent the medians of expected percentages across ensemble members, with uncertainty bars indicating the 5th to 95th percentile range. Blue bars indicate the observed percentages.



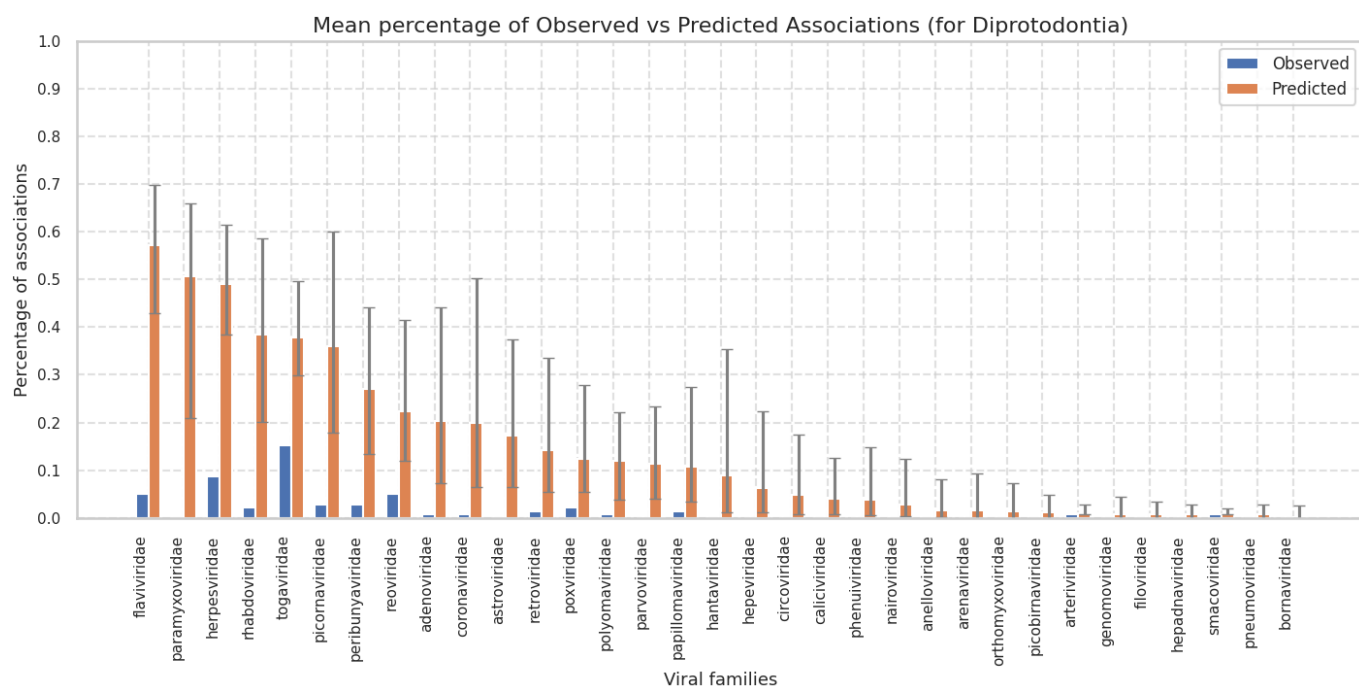
Supplementary Figure SR6 | Predicted associations by viral family for Dasyuromorphia. Percentage of predicted associations across viral families. Orange bars represent the medians of expected percentages across ensemble members, with uncertainty bars indicating the 5th to 95th percentile range. Blue bars indicate the observed percentages.



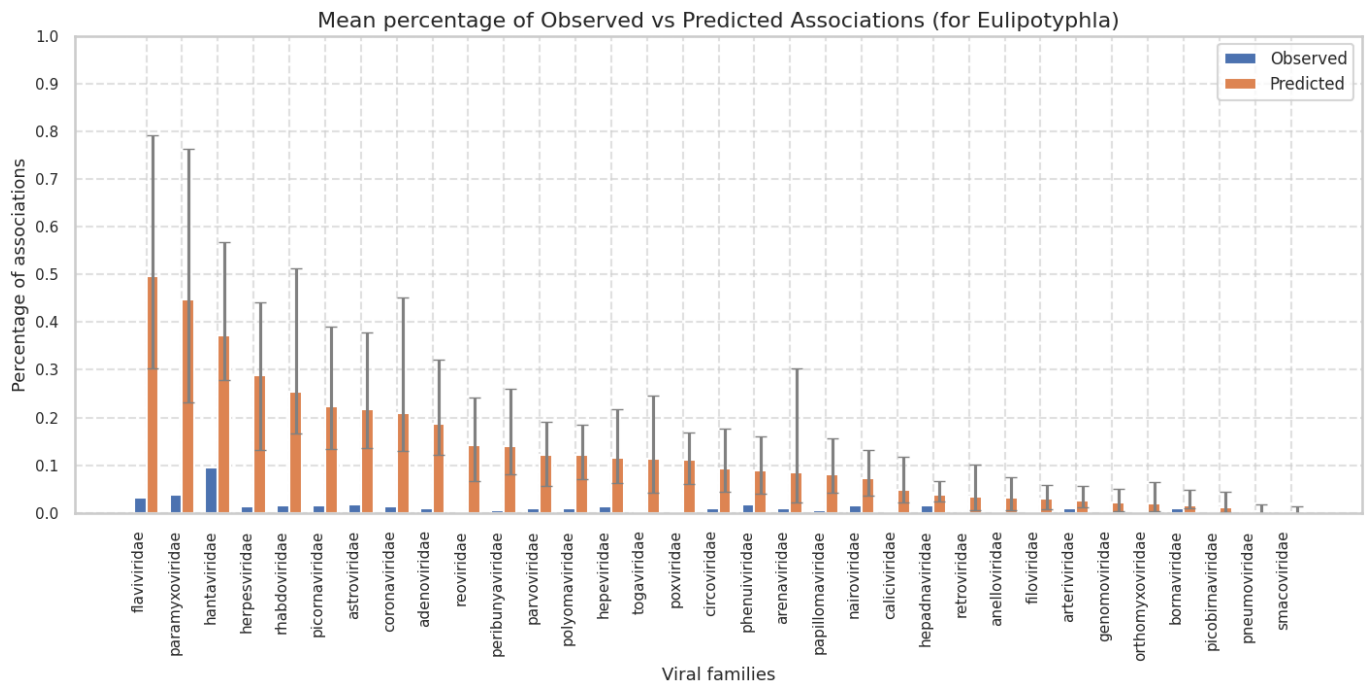
Supplementary Figure SR7 | Predicted associations by viral family for Dermoptera. Percentage of predicted associations across viral families. Orange bars represent the medians of expected percentages across ensemble members, with uncertainty bars indicating the 5th to 95th percentile range. Blue bars indicate the observed percentages.



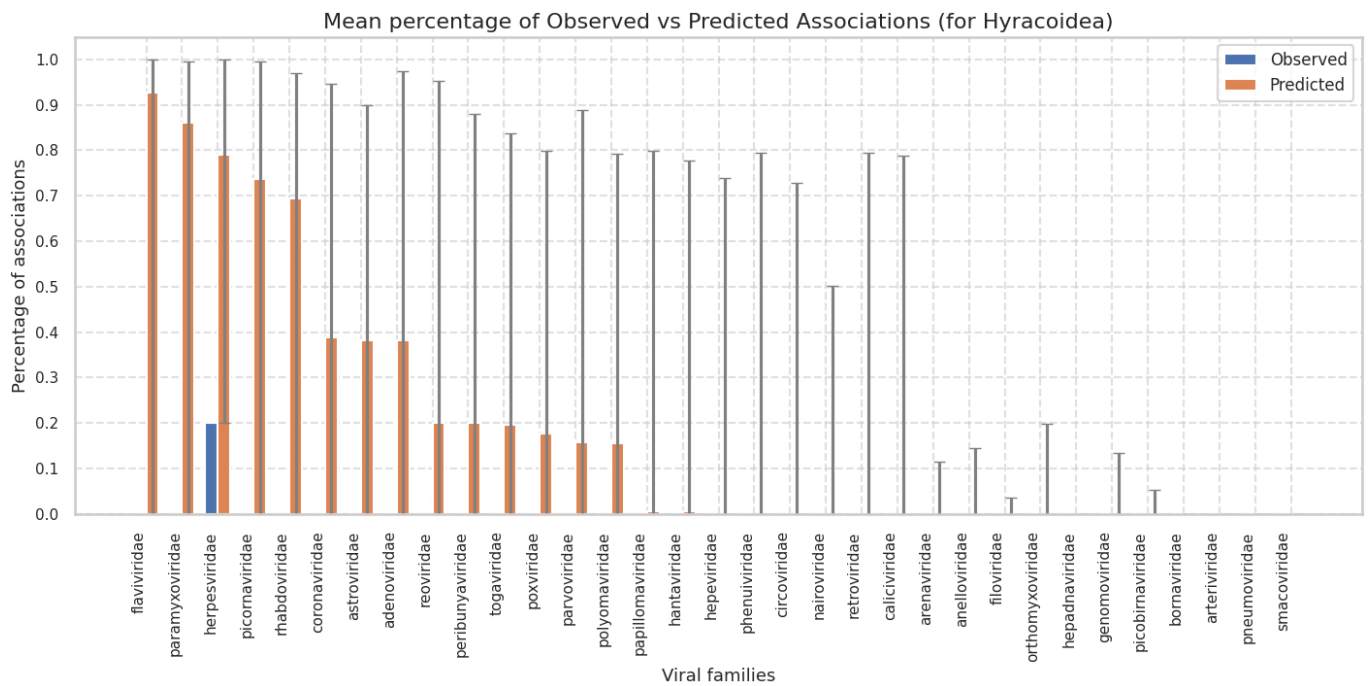
Supplementary Figure SR8 | Predicted associations by viral family for Didelphimorphia. Percentage of predicted associations across viral families. Orange bars represent the medians of expected percentages across ensemble members, with uncertainty bars indicating the 5th to 95th percentile range. Blue bars indicate the observed percentages.



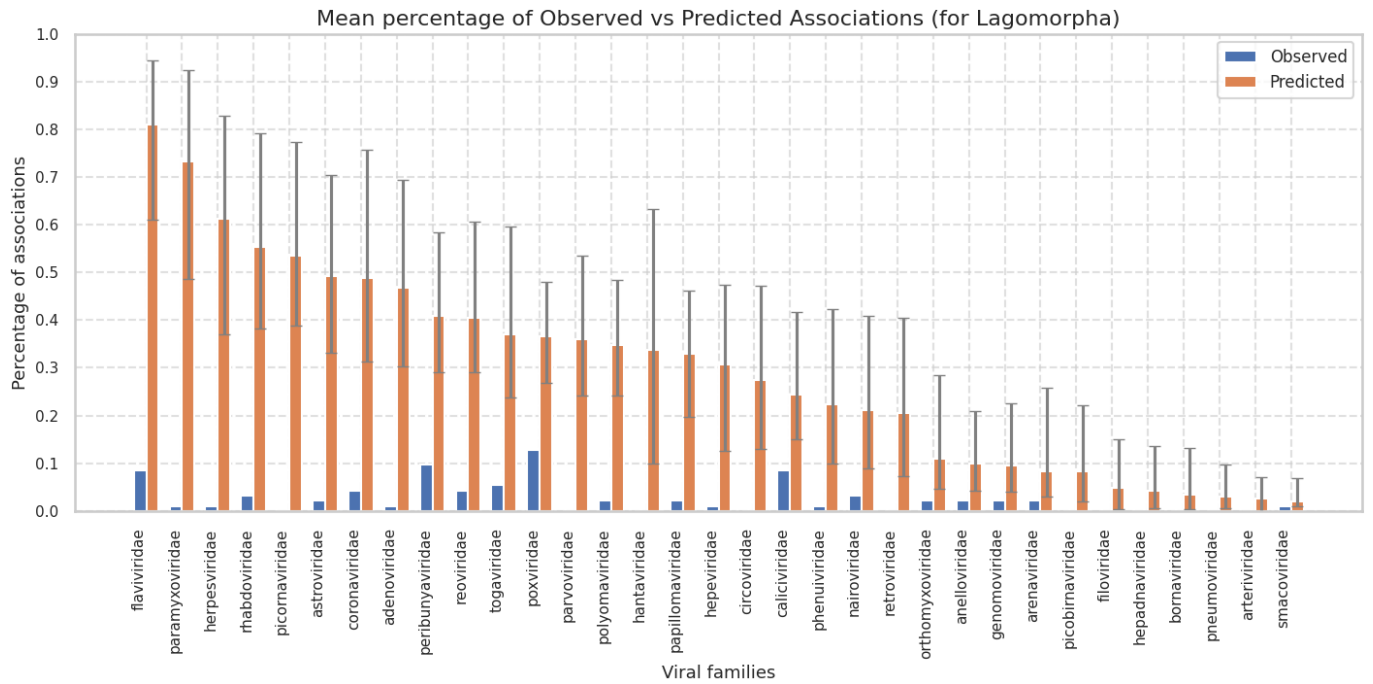
Supplementary Figure SR9 | Predicted associations by viral family for Diprotodontia. Percentage of predicted associations across viral families. Orange bars represent the medians of expected percentages across ensemble members, with uncertainty bars indicating the 5th to 95th percentile range. Blue bars indicate the observed percentages.



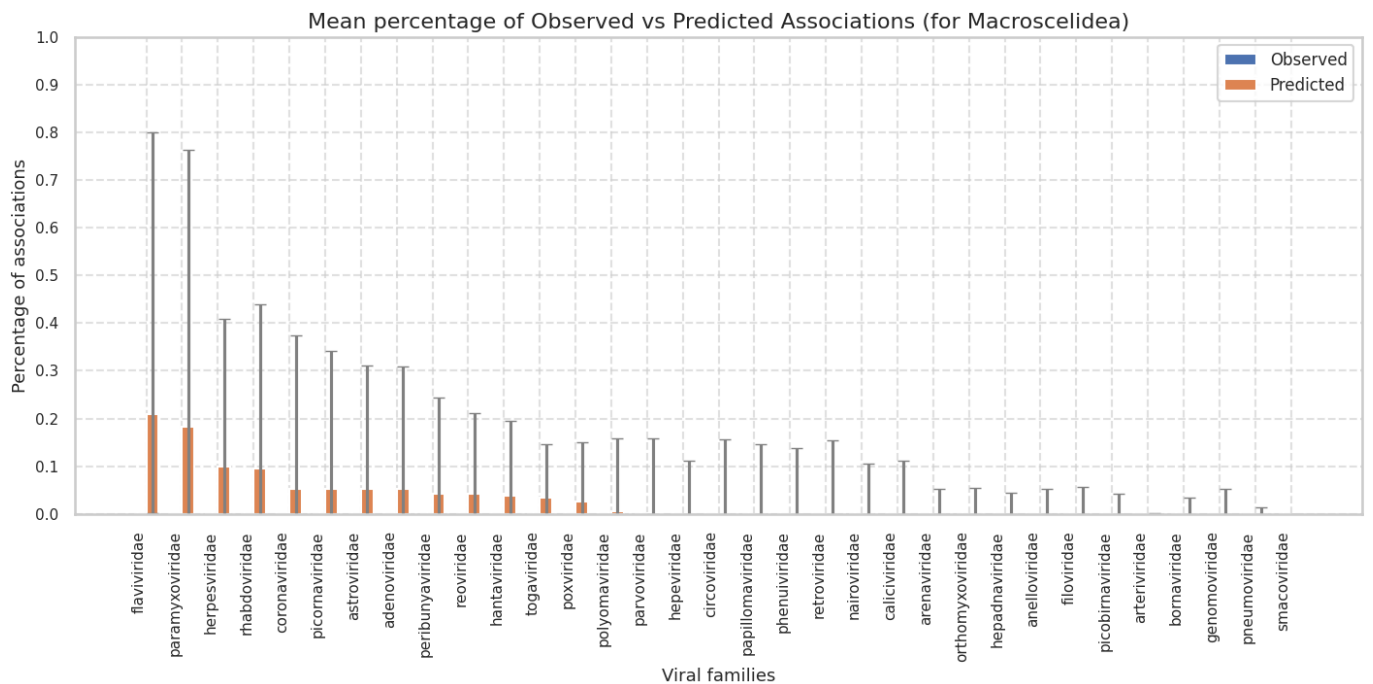
Supplementary Figure SR10 | Predicted associations by viral family for Eulipotyphla. Percentage of predicted associations across viral families. Orange bars represent the medians of expected percentages across ensemble members, with uncertainty bars indicating the 5th to 95th percentile range. Blue bars indicate the observed percentages.



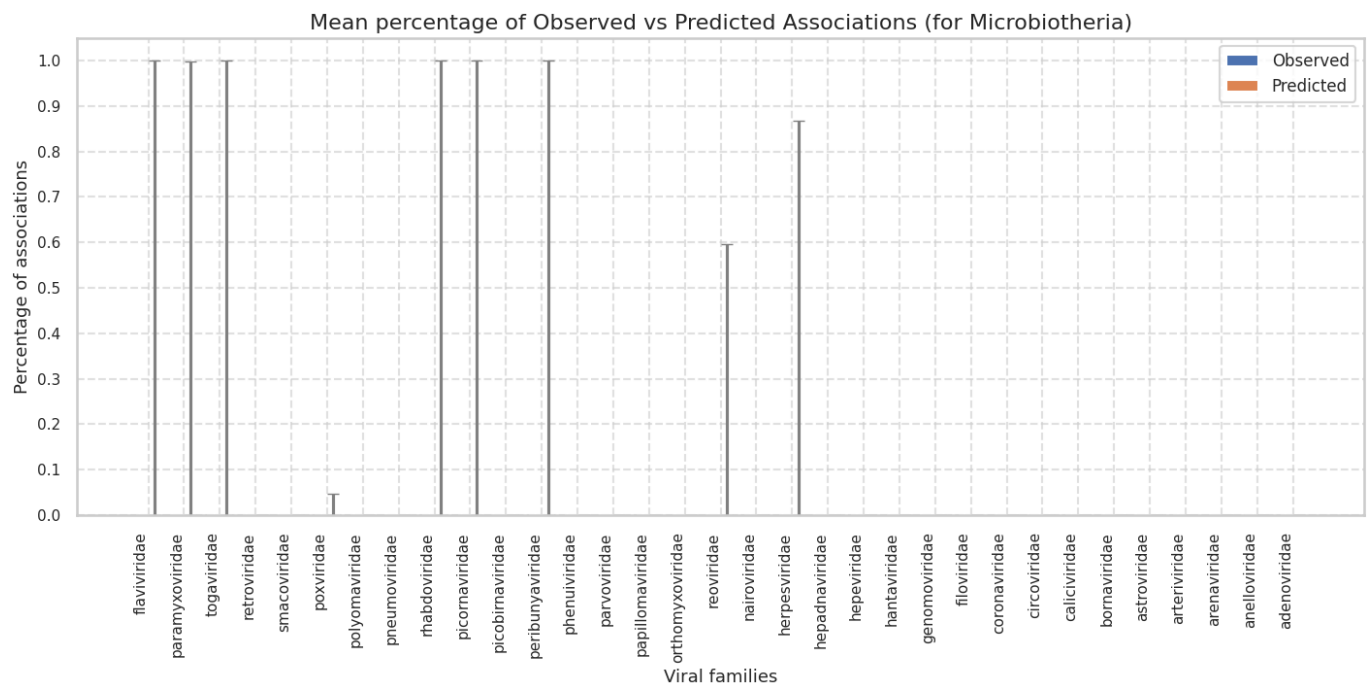
Supplementary Figure SR11 | Predicted associations by viral family for Hyracoidea. Percentage of predicted associations across viral families. Orange bars represent the medians of expected percentages across ensemble members, with uncertainty bars indicating the 5th to 95th percentile range. Blue bars indicate the observed percentages.



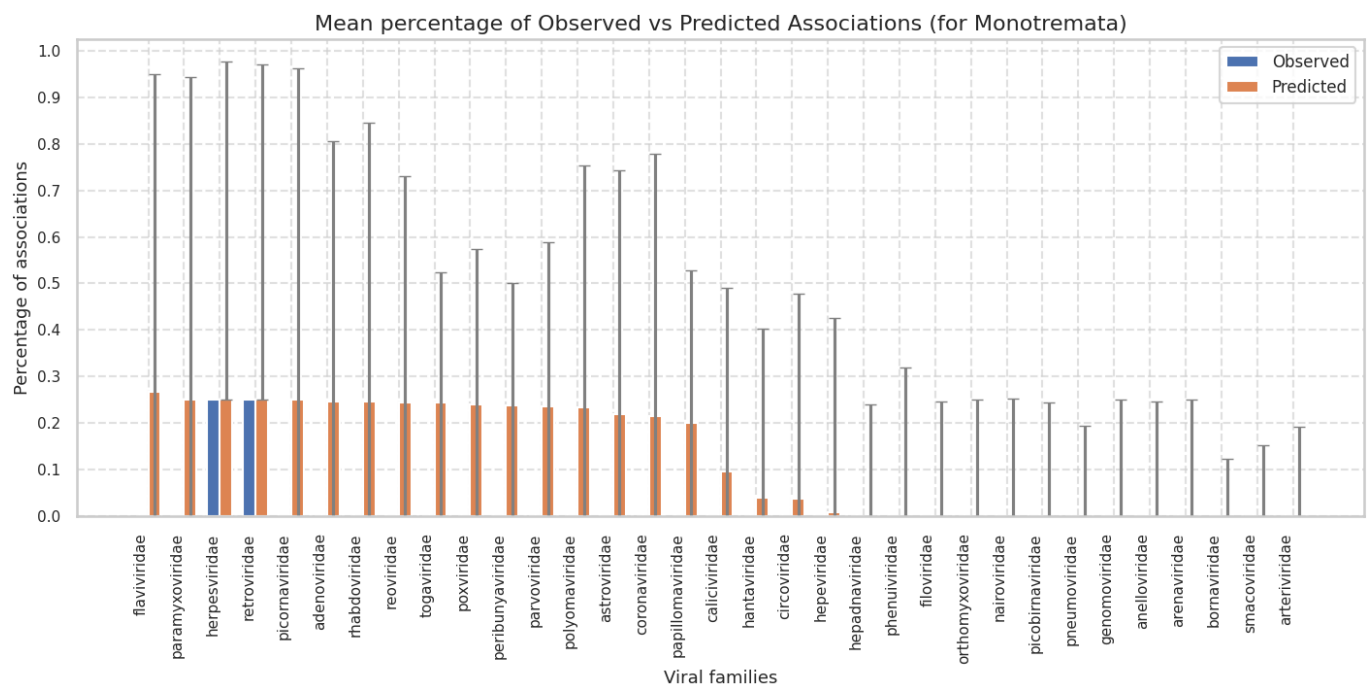
Supplementary Figure SR12 | Predicted associations by viral family for Lagomorpha. Percentage of predicted associations across viral families. Orange bars represent the medians of expected percentages across ensemble members, with uncertainty bars indicating the 5th to 95th percentile range. Blue bars indicate the observed percentages.



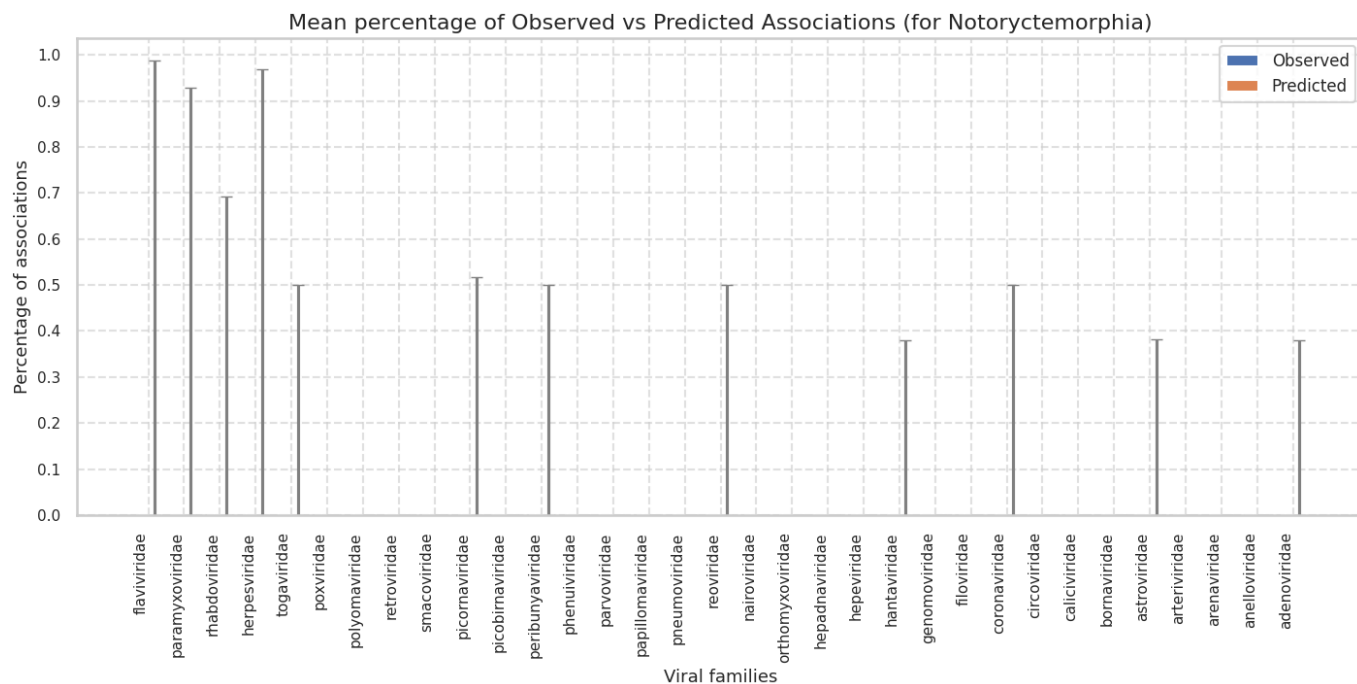
Supplementary Figure SR13 | Predicted associations by viral family for Macroscelidea. Percentage of predicted associations across viral families. Orange bars represent the medians of expected percentages across ensemble members, with uncertainty bars indicating the 5th to 95th percentile range. Blue bars indicate the observed percentages.



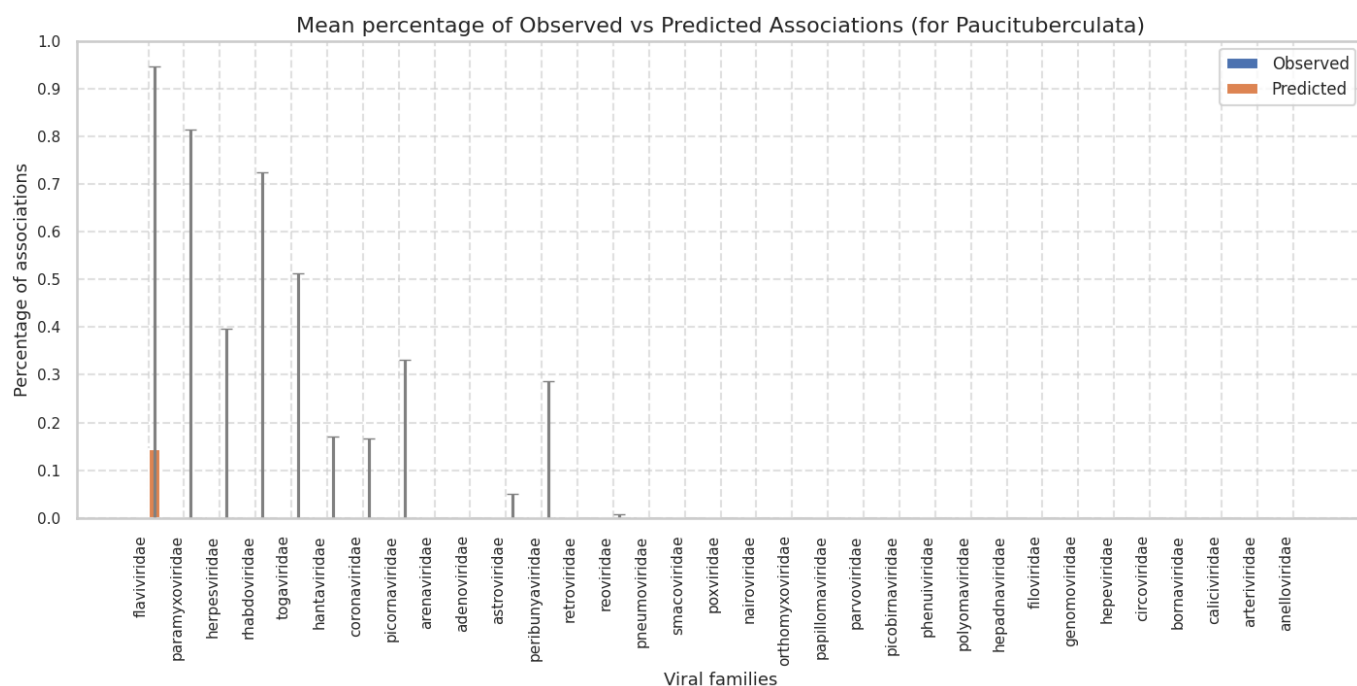
Supplementary Figure SR14 | Predicted associations by viral family for Microbiotheria. Percentage of predicted associations across viral families. Orange bars represent the medians of expected percentages across ensemble members, with uncertainty bars indicating the 5th to 95th percentile range. Blue bars indicate the observed percentages.



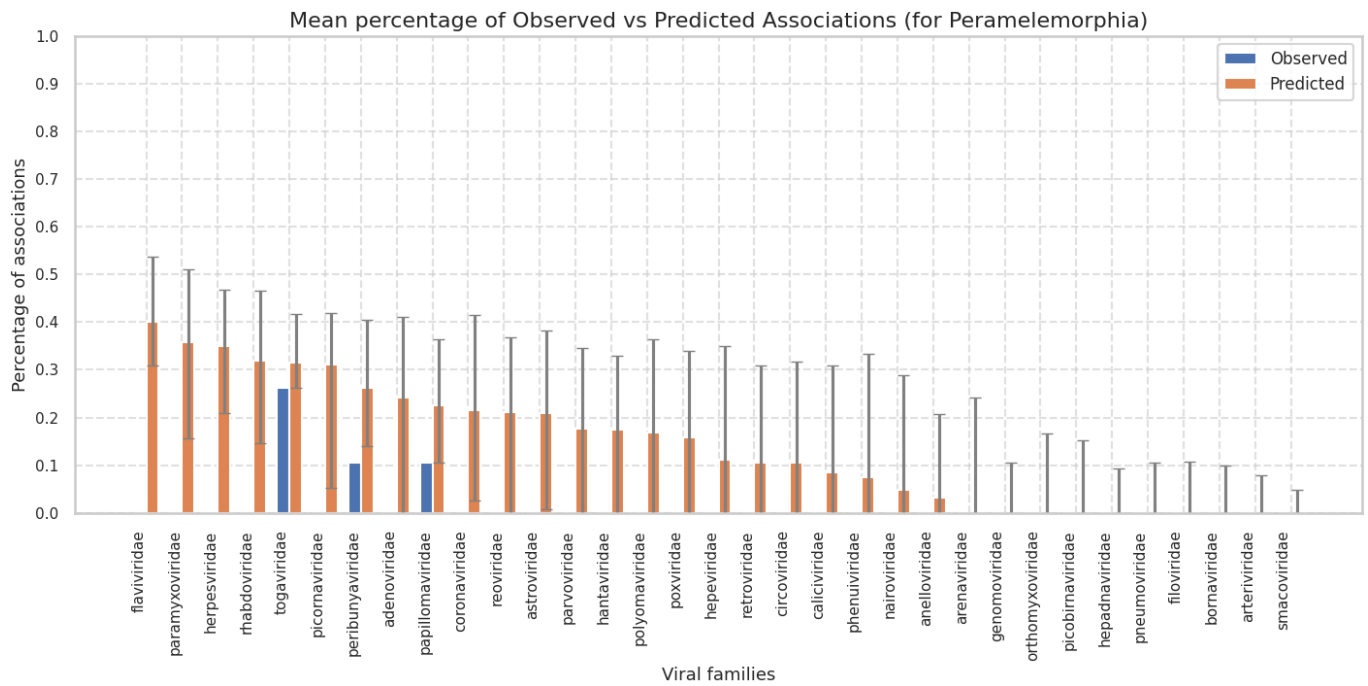
Supplementary Figure SR15 | Predicted associations by viral family for Monotremata. Percentage of predicted associations across viral families. Orange bars represent the medians of expected percentages across ensemble members, with uncertainty bars indicating the 5th to 95th percentile range. Blue bars indicate the observed percentages.



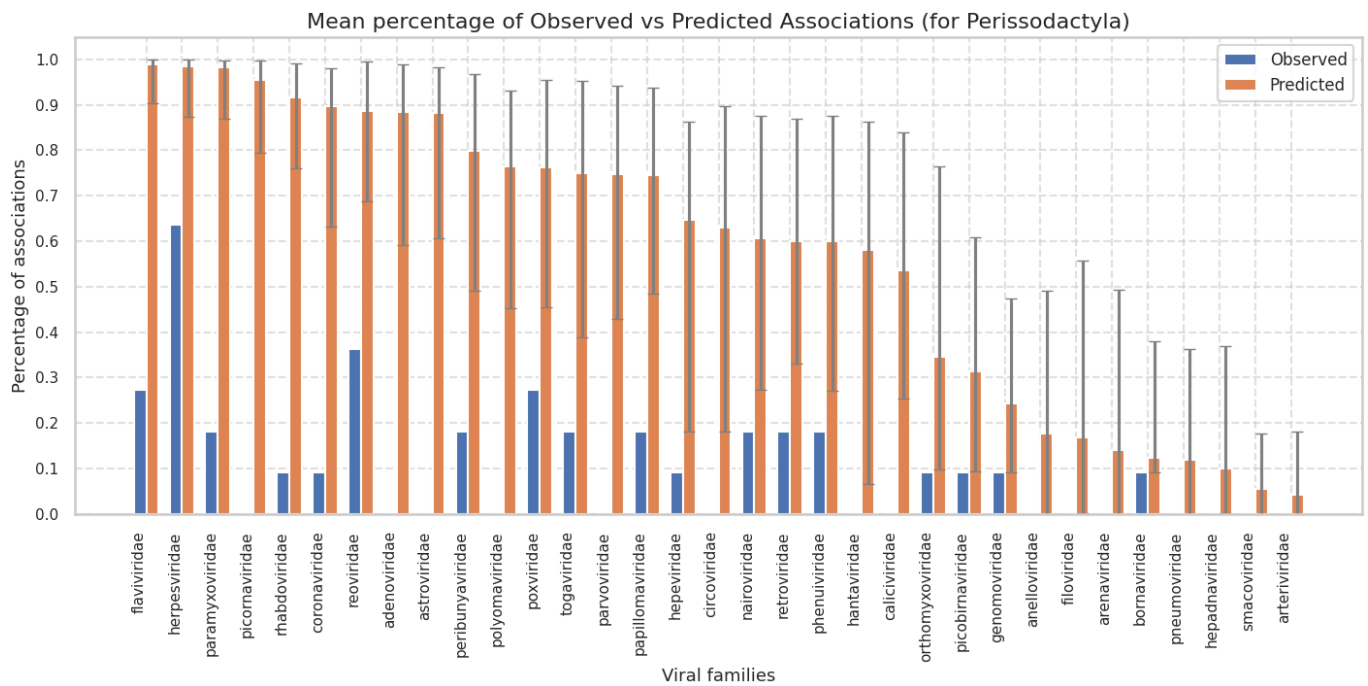
Supplementary Figure SR16 | Predicted associations by viral family for Notoryctemorphia. Percentage of predicted associations across viral families. Orange bars represent the medians of expected percentages across ensemble members, with uncertainty bars indicating the 5th to 95th percentile range. Blue bars indicate the observed percentages.



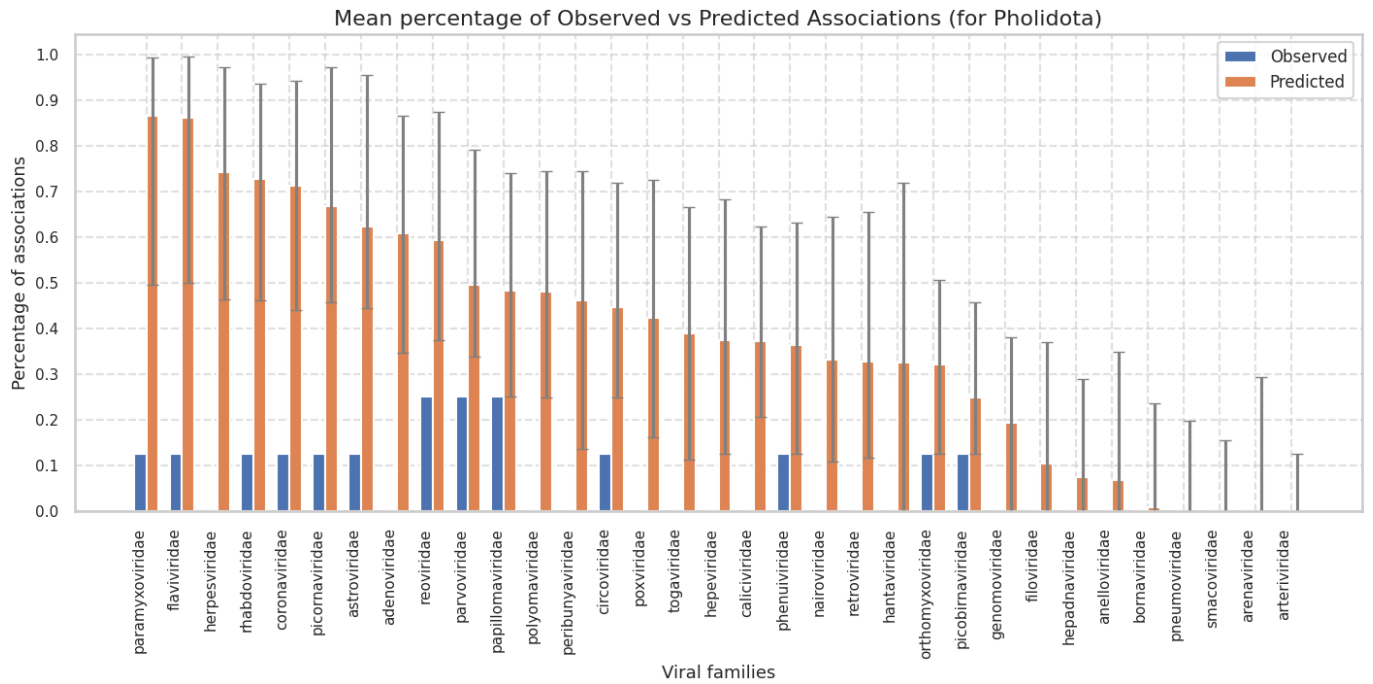
Supplementary Figure SR17 | Predicted associations by viral family for Paucituberculata. Percentage of predicted associations across viral families. Orange bars represent the medians of expected percentages across ensemble members, with uncertainty bars indicating the 5th to 95th percentile range. Blue bars indicate the observed percentages.



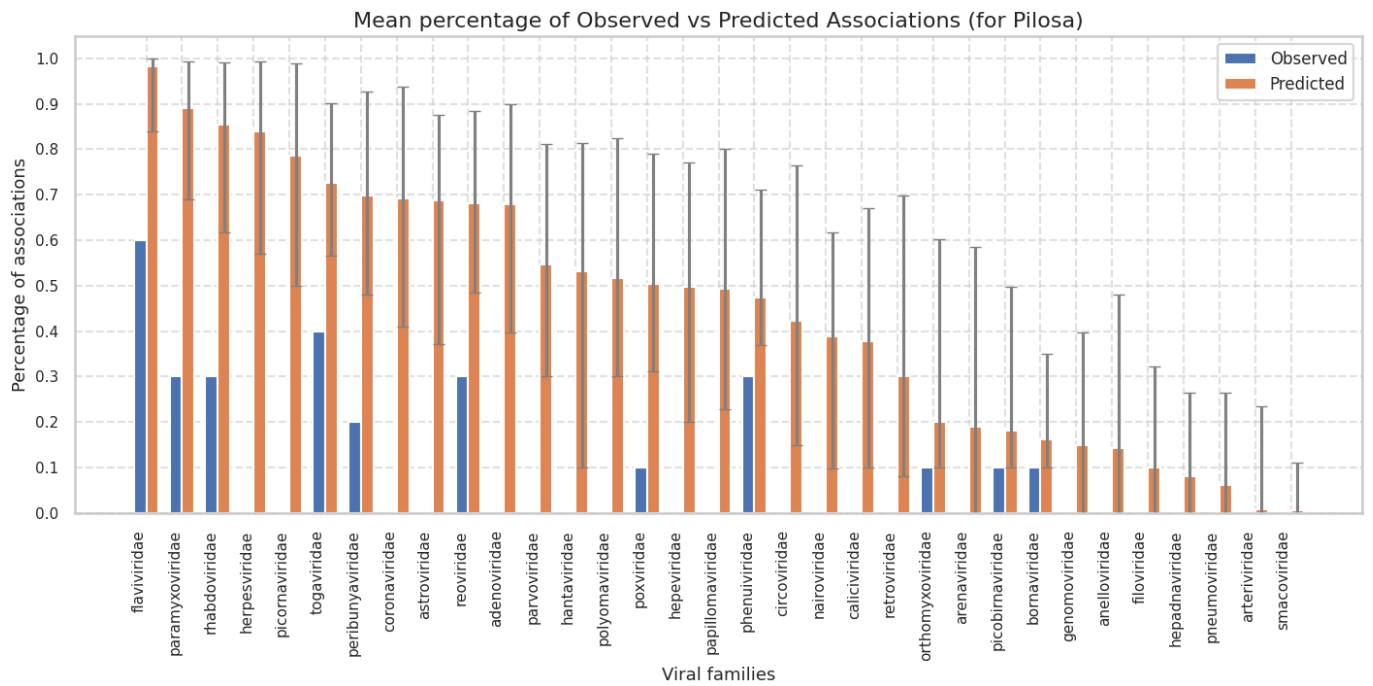
Supplementary Figure SR18 | Predicted associations by viral family for Peramelemorphia. Percentage of predicted associations across viral families. Orange bars represent the medians of expected percentages across ensemble members, with uncertainty bars indicating the 5th to 95th percentile range. Blue bars indicate the observed percentages.



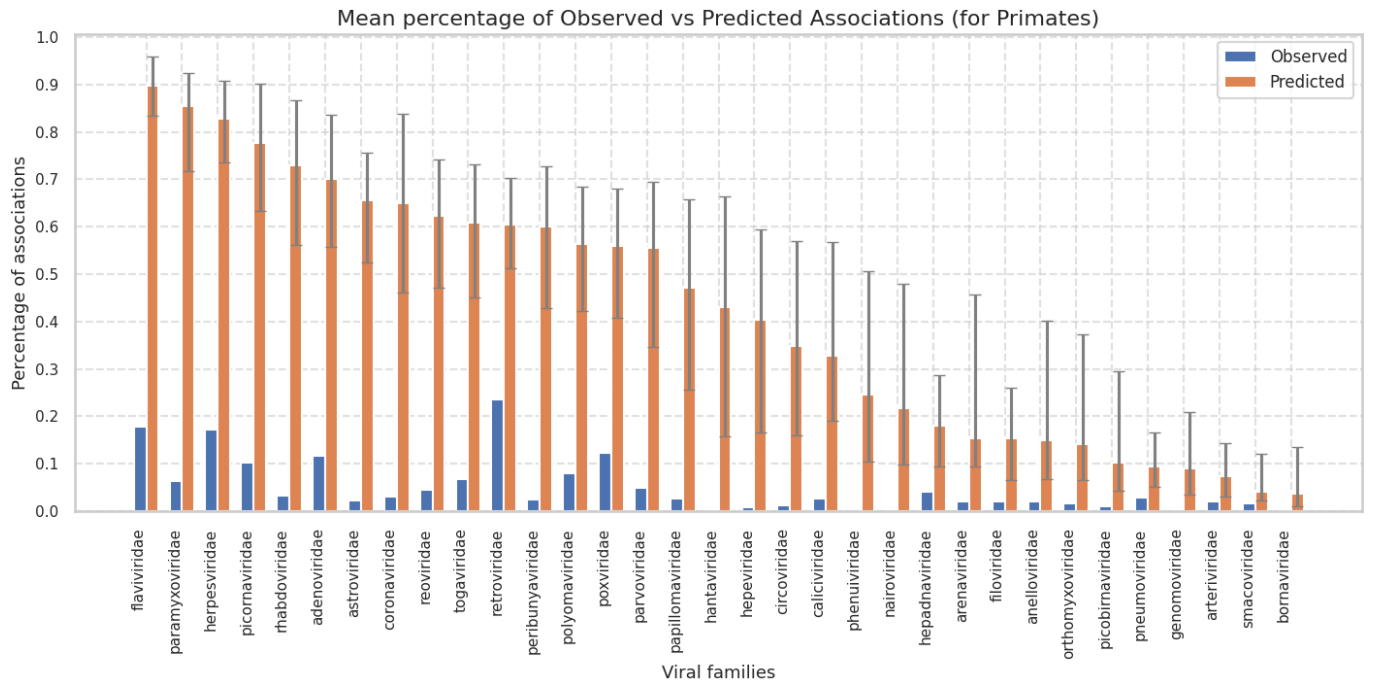
Supplementary Figure SR19 | Predicted associations by viral family for Perissodactyla. Percentage of predicted associations across viral families. Orange bars represent the medians of expected percentages across ensemble members, with uncertainty bars indicating the 5th to 95th percentile range. Blue bars indicate the observed percentages.



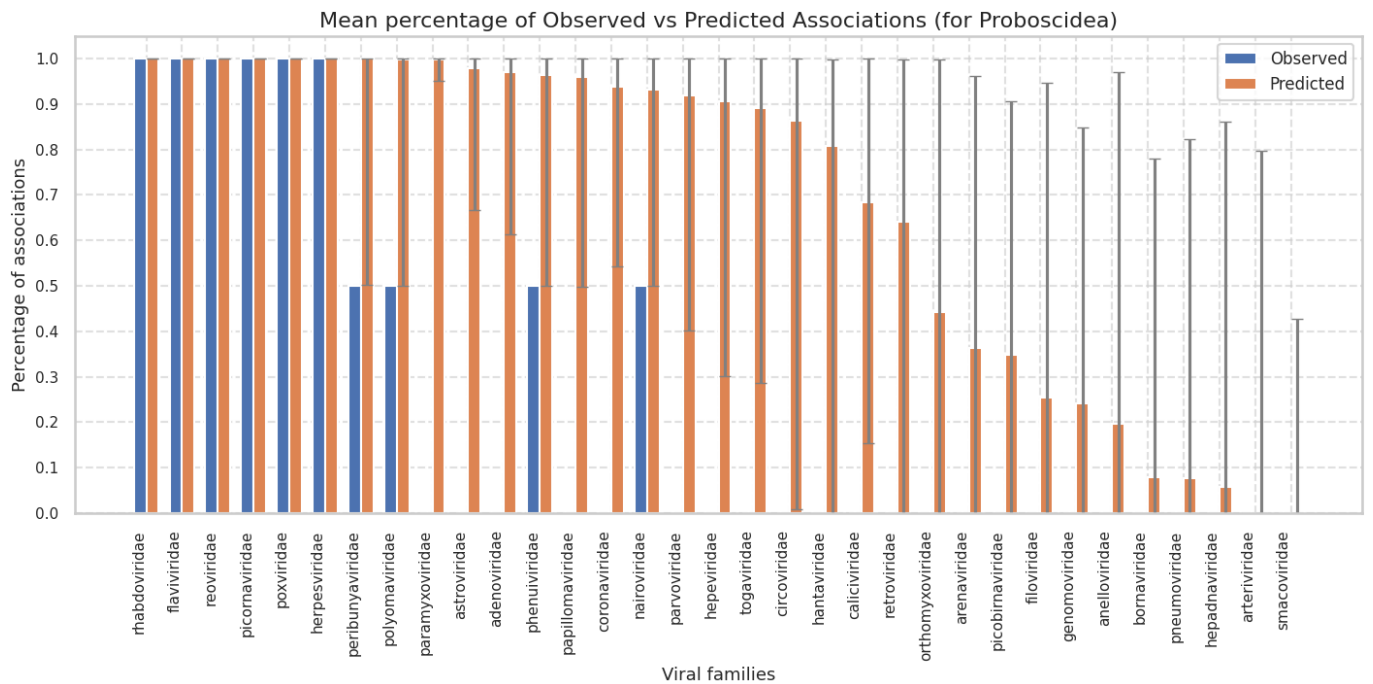
Supplementary Figure SR20 | Predicted associations by viral family for Pholidota. Percentage of predicted associations across viral families. Orange bars represent the medians of expected percentages across ensemble members, with uncertainty bars indicating the 5th to 95th percentile range. Blue bars indicate the observed percentages.



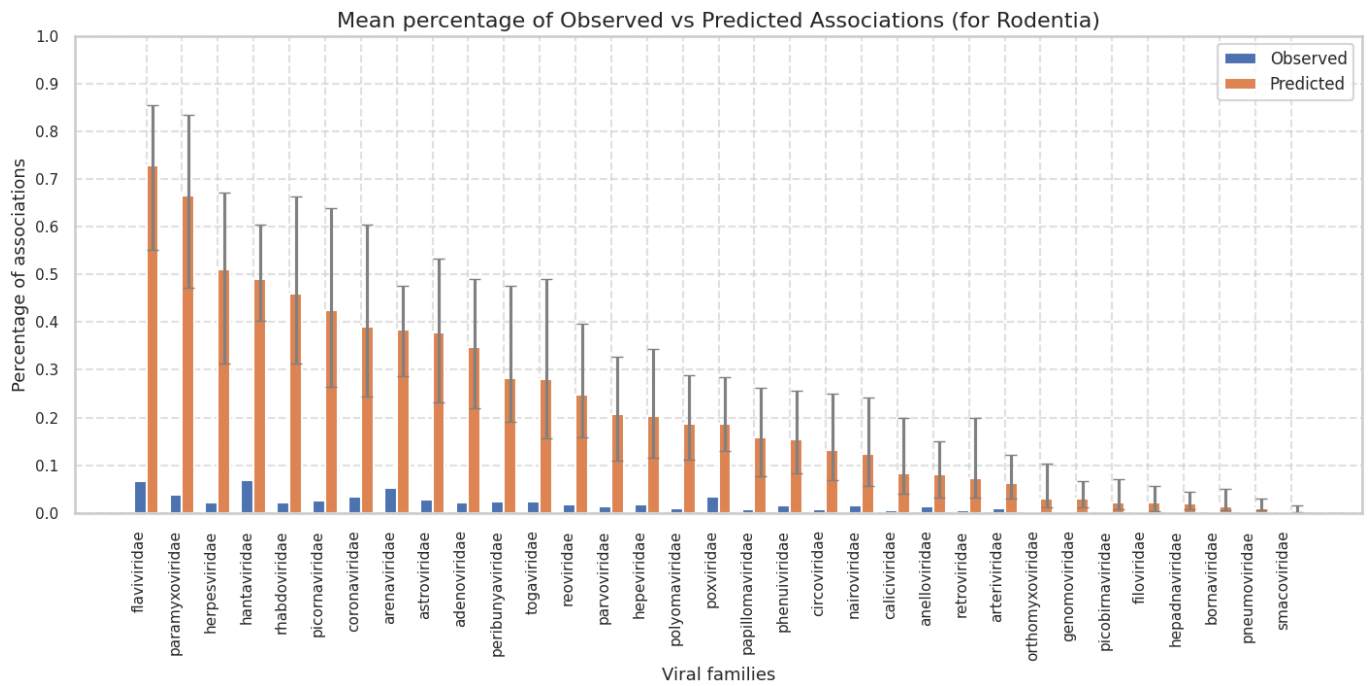
Supplementary Figure SR21 | Predicted associations by viral family for Pilosa. Percentage of predicted associations across viral families. Orange bars represent the medians of expected percentages across ensemble members, with uncertainty bars indicating the 5th to 95th percentile range. Blue bars indicate the observed percentages.



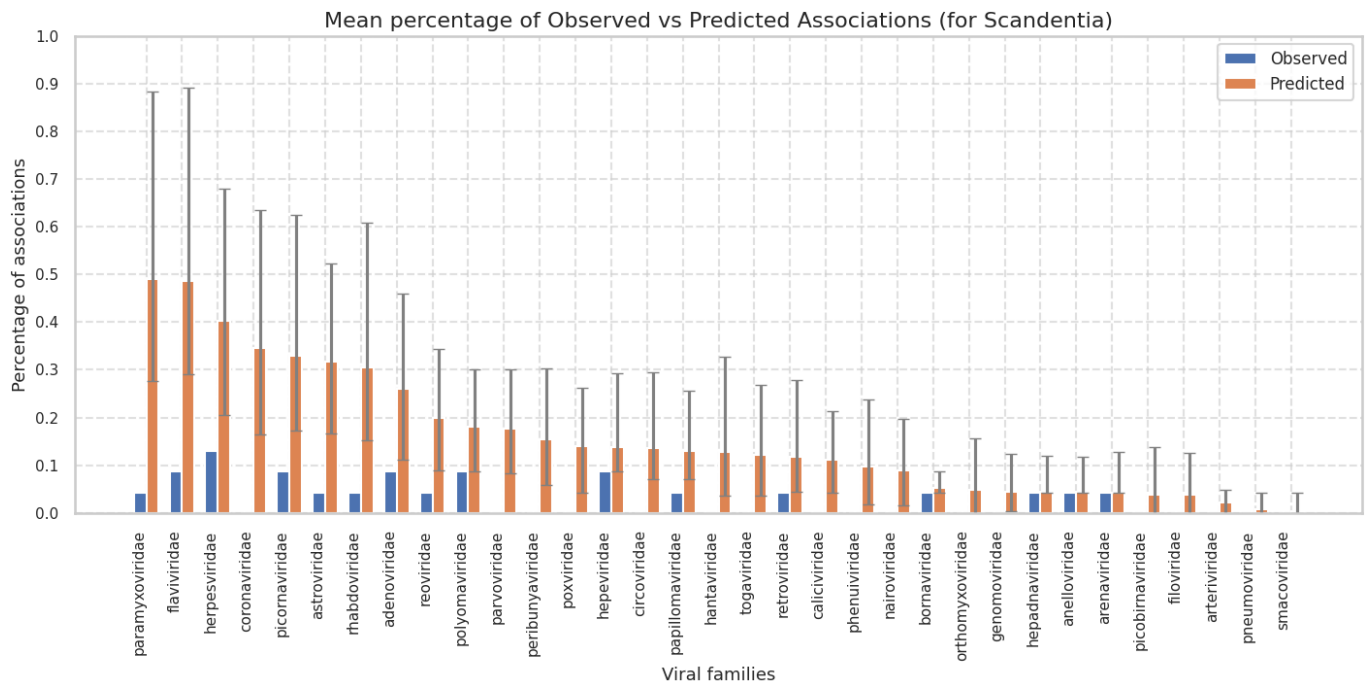
Supplementary Figure SR22 | Predicted associations by viral family for Primates. Percentage of predicted associations across viral families. Orange bars represent the medians of expected percentages across ensemble members, with uncertainty bars indicating the 5th to 95th percentile range. Blue bars indicate the observed percentages.



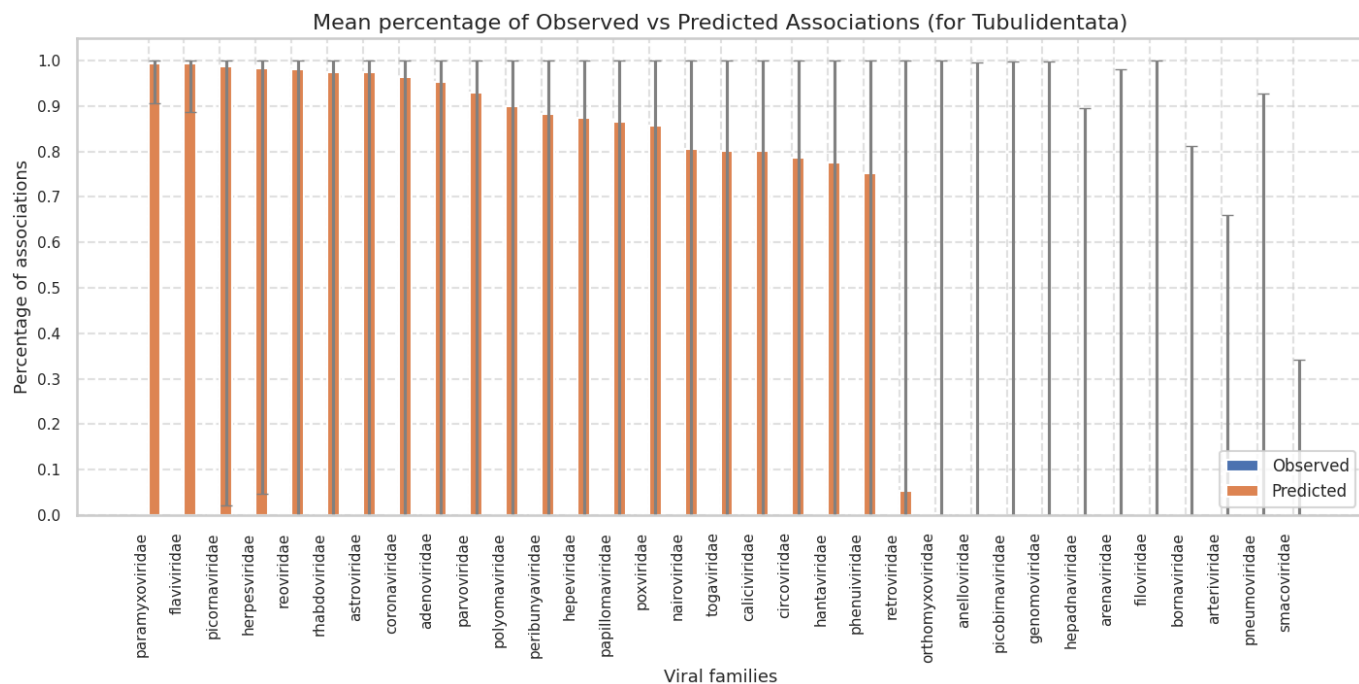
Supplementary Figure SR23 | Predicted associations by viral family for Proboscidea. Percentage of predicted associations across viral families. Orange bars represent the medians of expected percentages across ensemble members, with uncertainty bars indicating the 5th to 95th percentile range. Blue bars indicate the observed percentages.



Supplementary Figure SR24 | Predicted associations by viral family for Rodentia. Percentage of predicted associations across viral families. Orange bars represent the medians of expected percentages across ensemble members, with uncertainty bars indicating the 5th to 95th percentile range. Blue bars indicate the observed percentages.



Supplementary Figure SR25 | Predicted associations by viral family for Scandentia. Percentage of predicted associations across viral families. Orange bars represent the medians of expected percentages across ensemble members, with uncertainty bars indicating the 5th to 95th percentile range. Blue bars indicate the observed percentages.

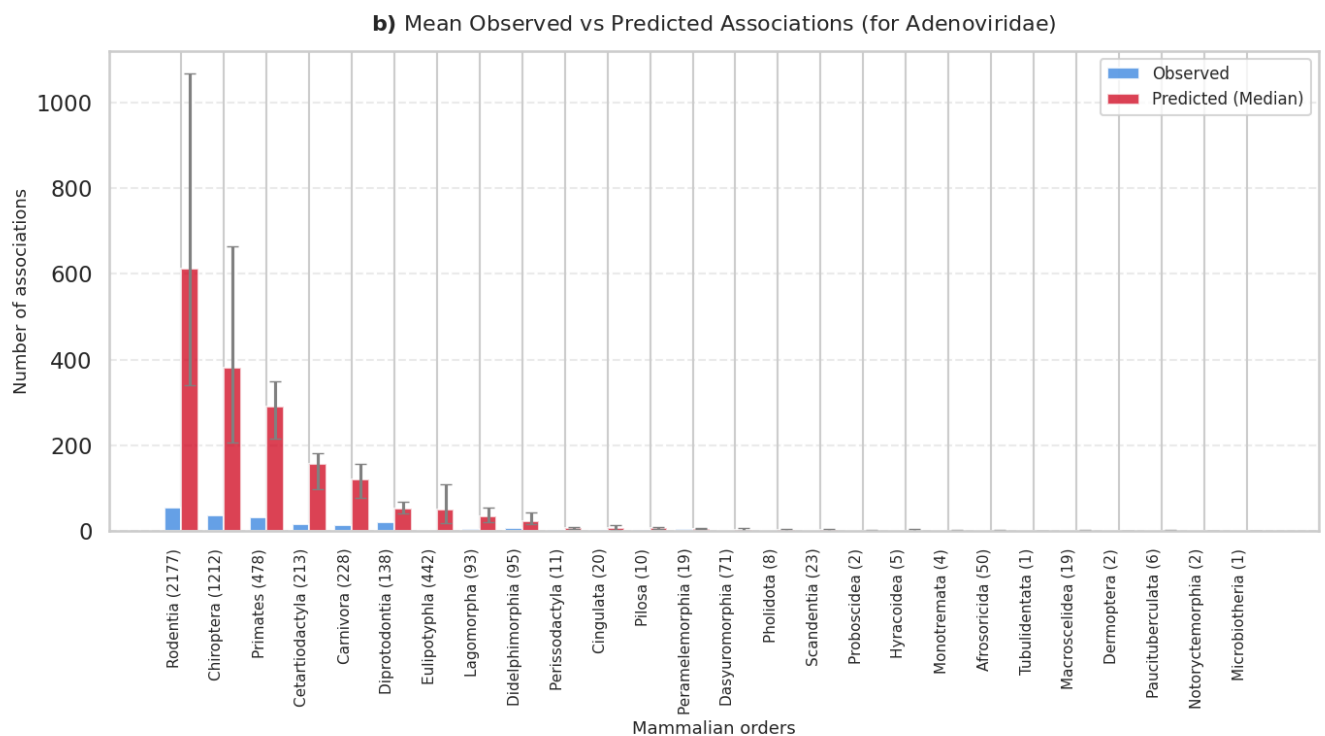
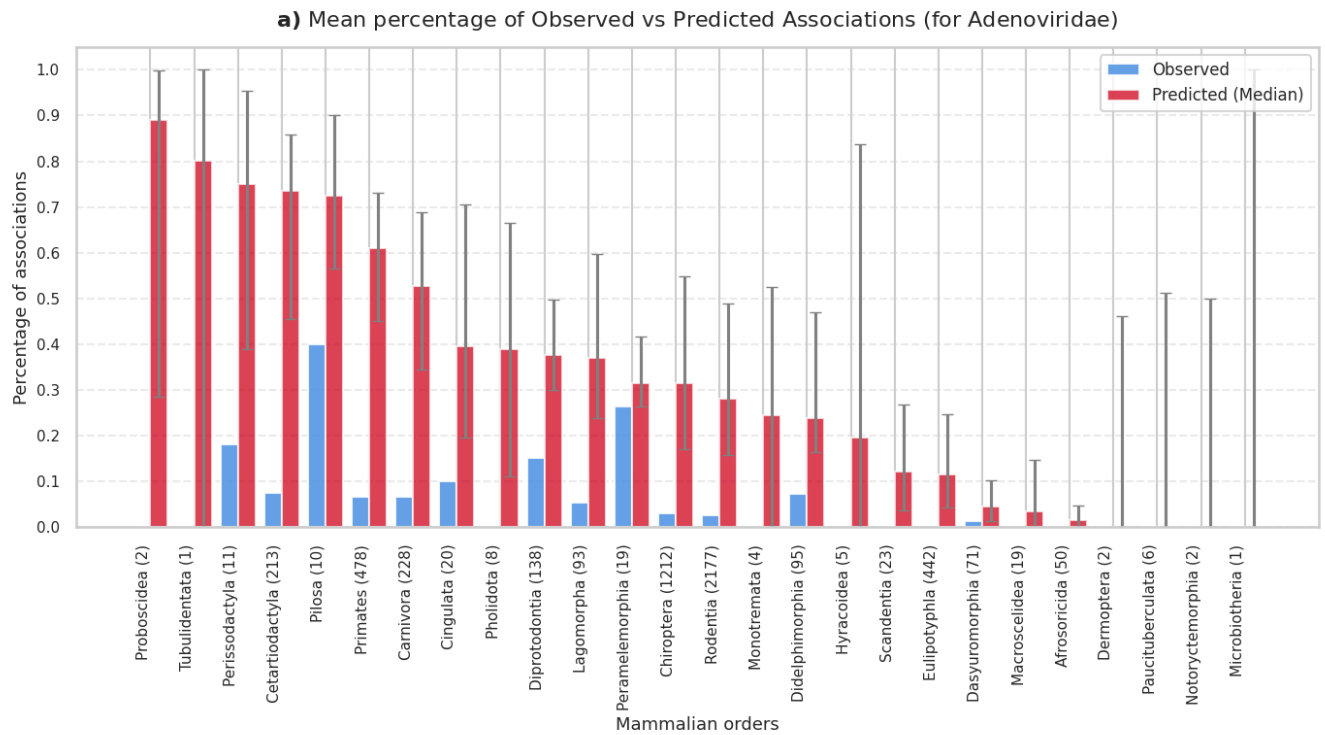


Supplementary Figure SR26 | Predicted associations by viral family for Tubulidentata. Percentage of predicted associations across viral families. Orange bars represent the medians of expected percentages across ensemble members, with uncertainty bars indicating the 5th to 95th percentile range. Blue bars indicate the observed percentages.

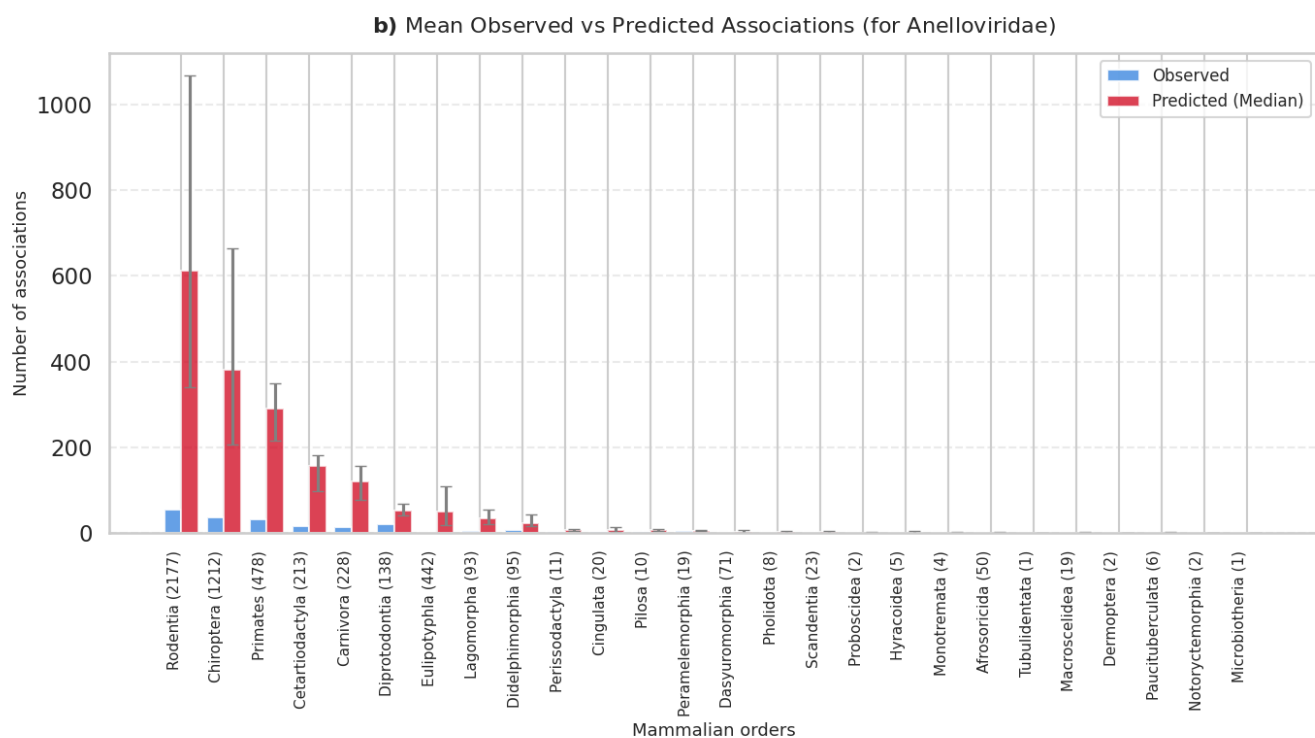
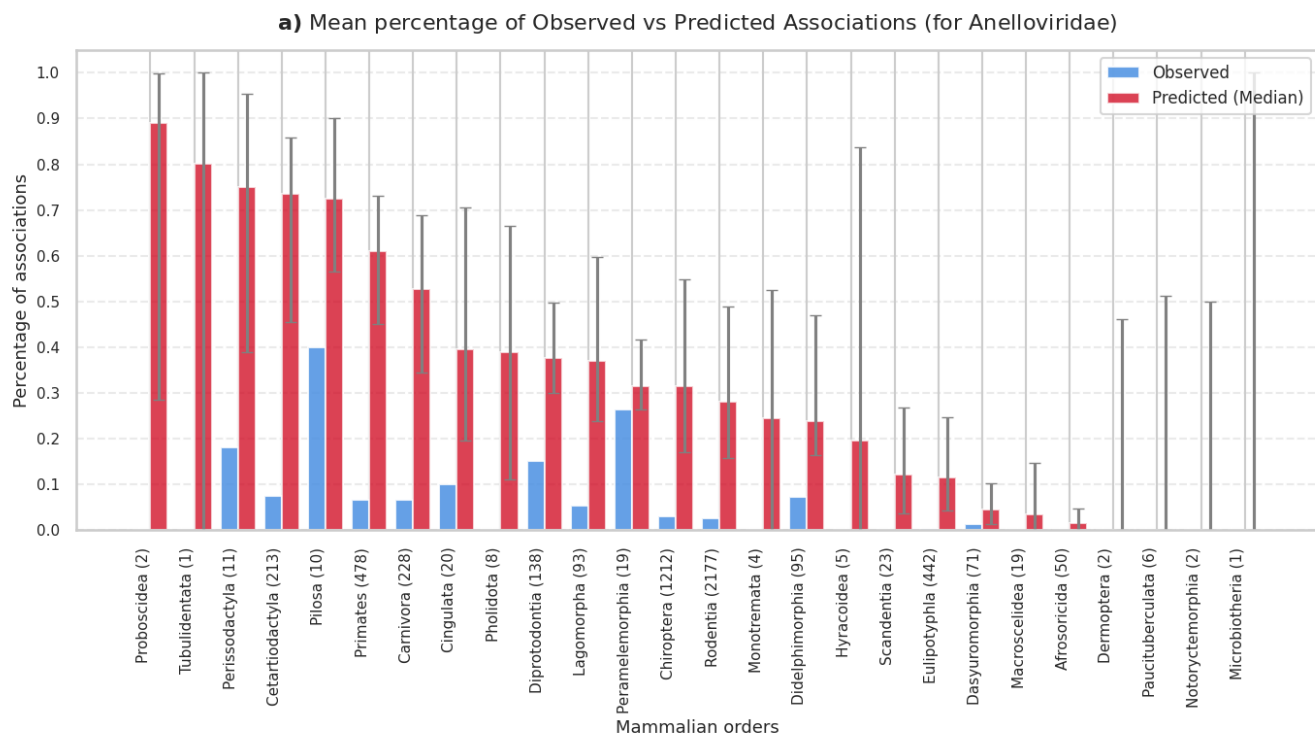
2 Predictions for each viral families

2.1 Predictions across mammalian orders

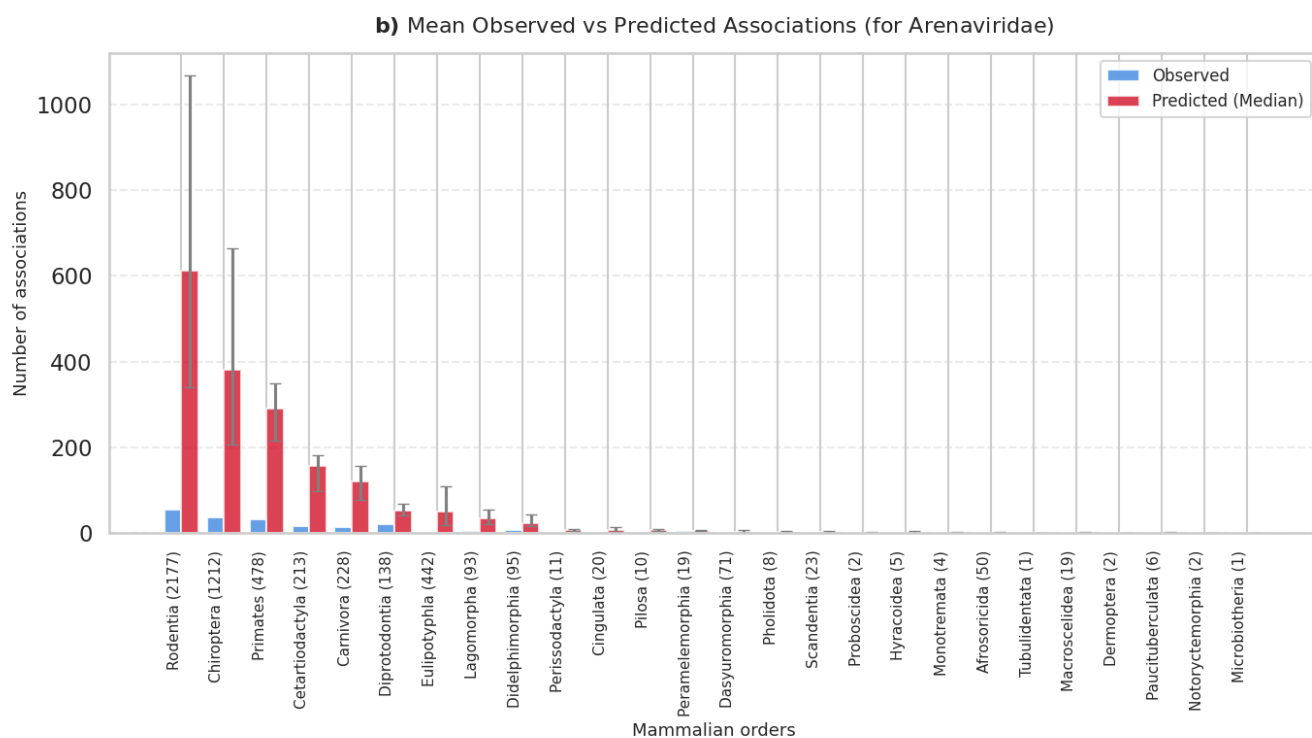
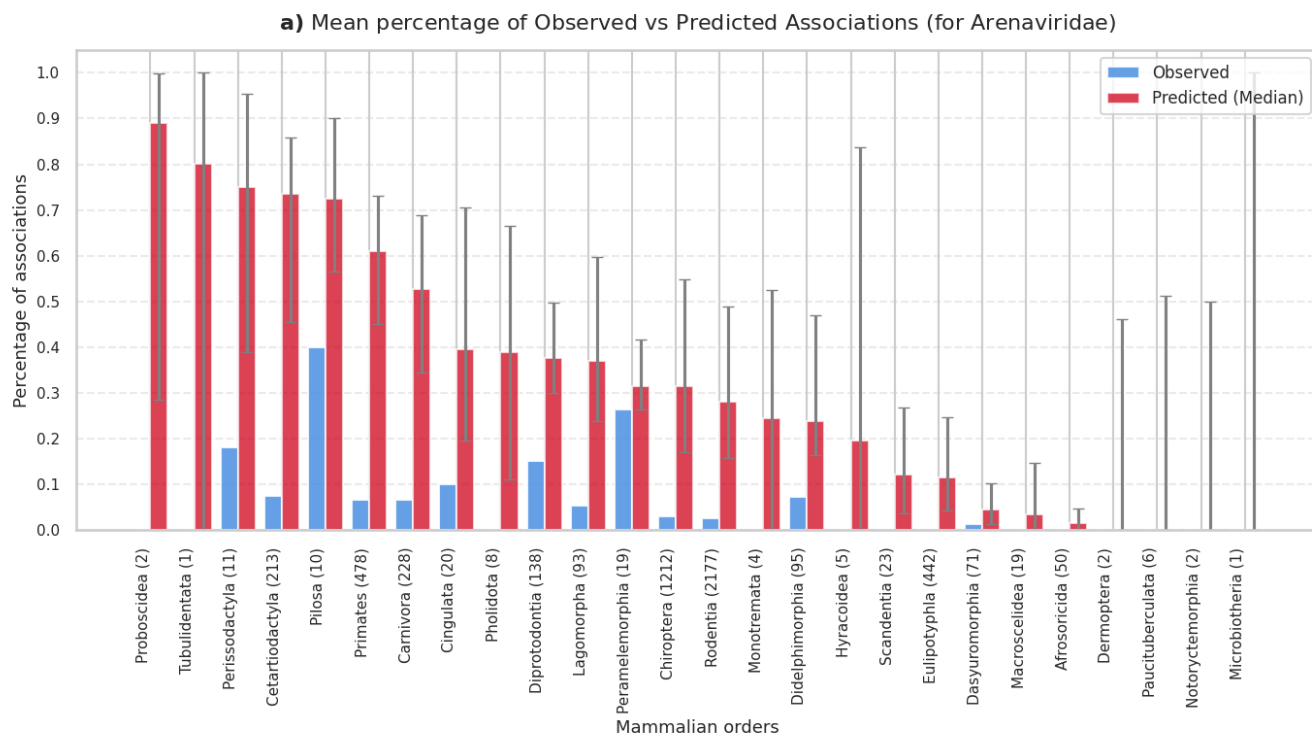
Here, for each viral family, we show the number and the percentage of observed vs predicted associations across mammalian orders (Supplementary Figures SR27–SR59).



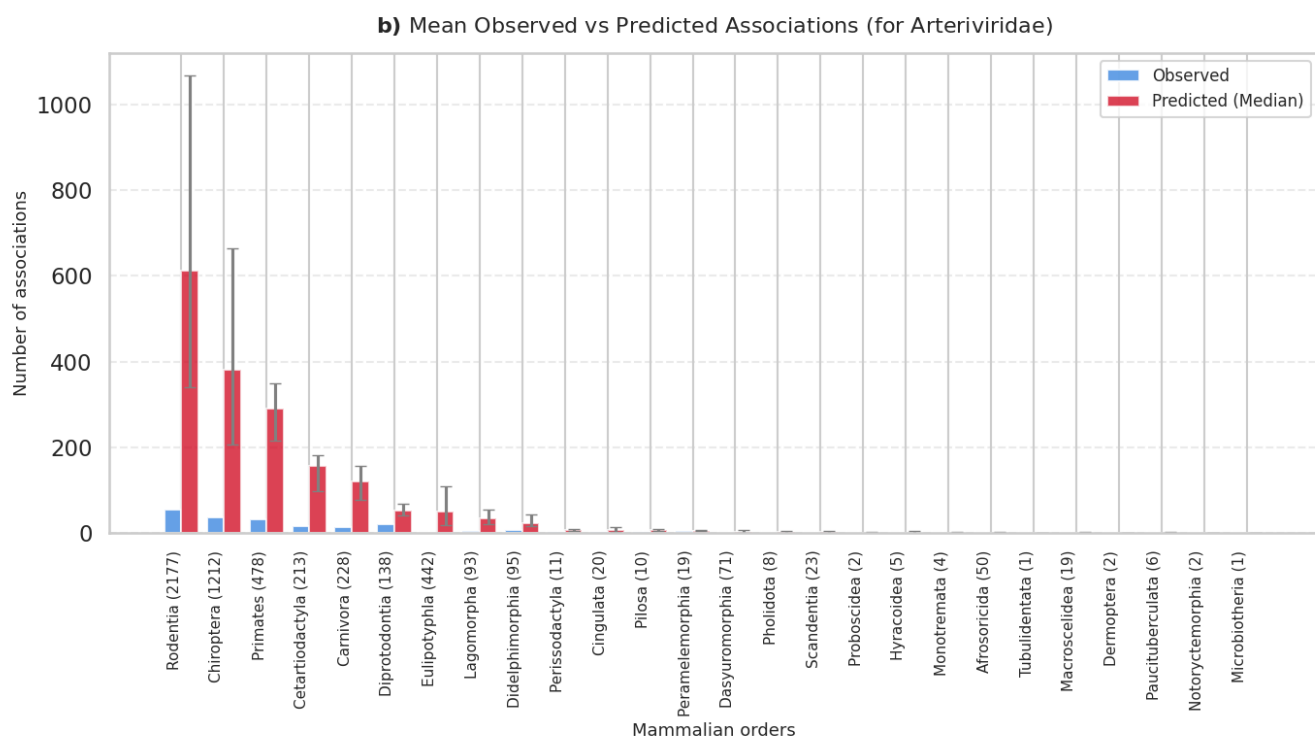
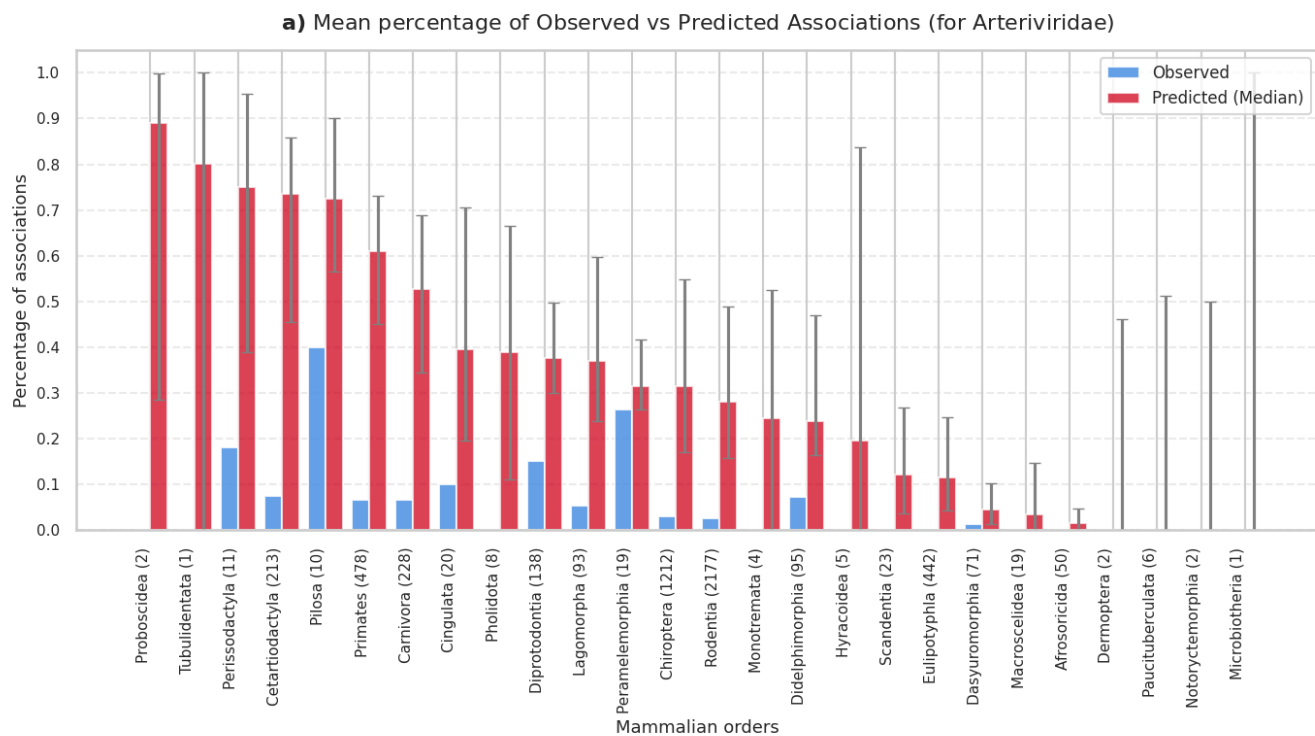
Supplementary Figure SR27 | Predicted associations by mammalian order for Adenoviridae. Percentage (a) and total number (b) of predicted associations across mammalian orders. Orange bars represent the medians of expected percentages across ensemble members, with uncertainty bars indicating the 5th to 95th percentile range. Blue bars indicate the observed percentages.



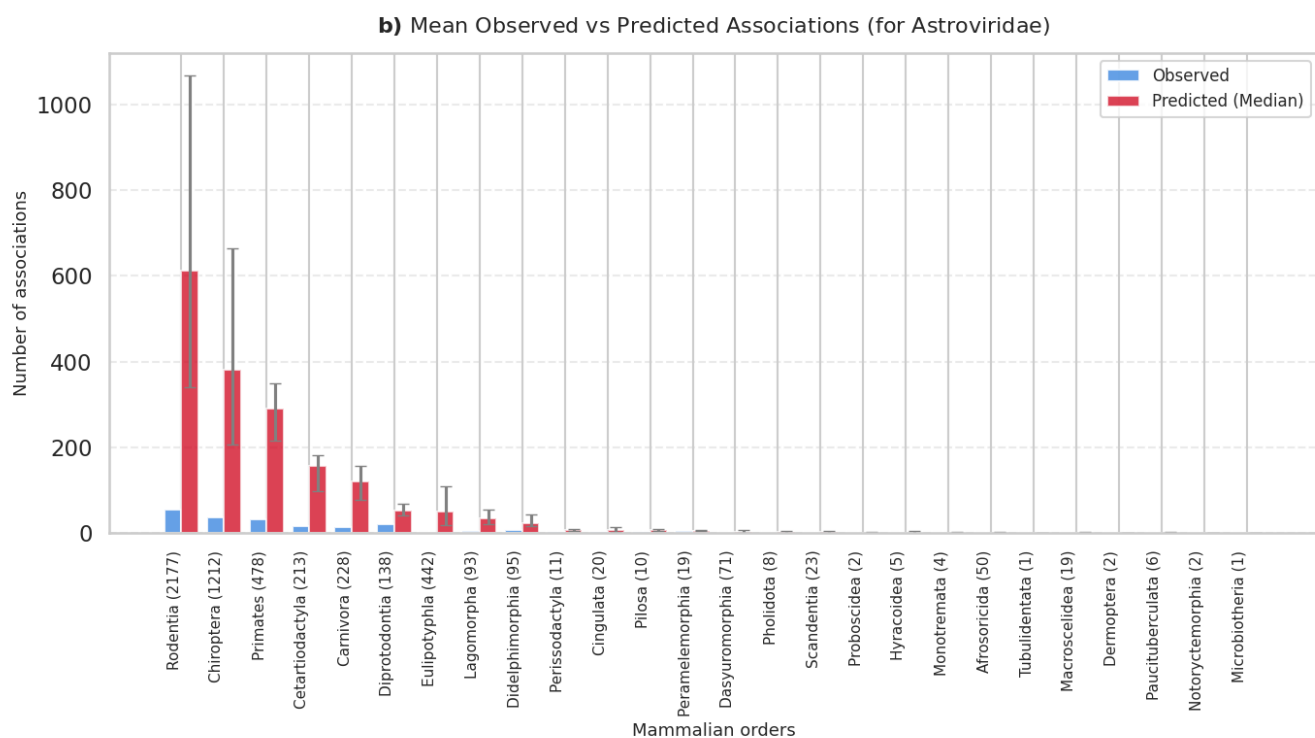
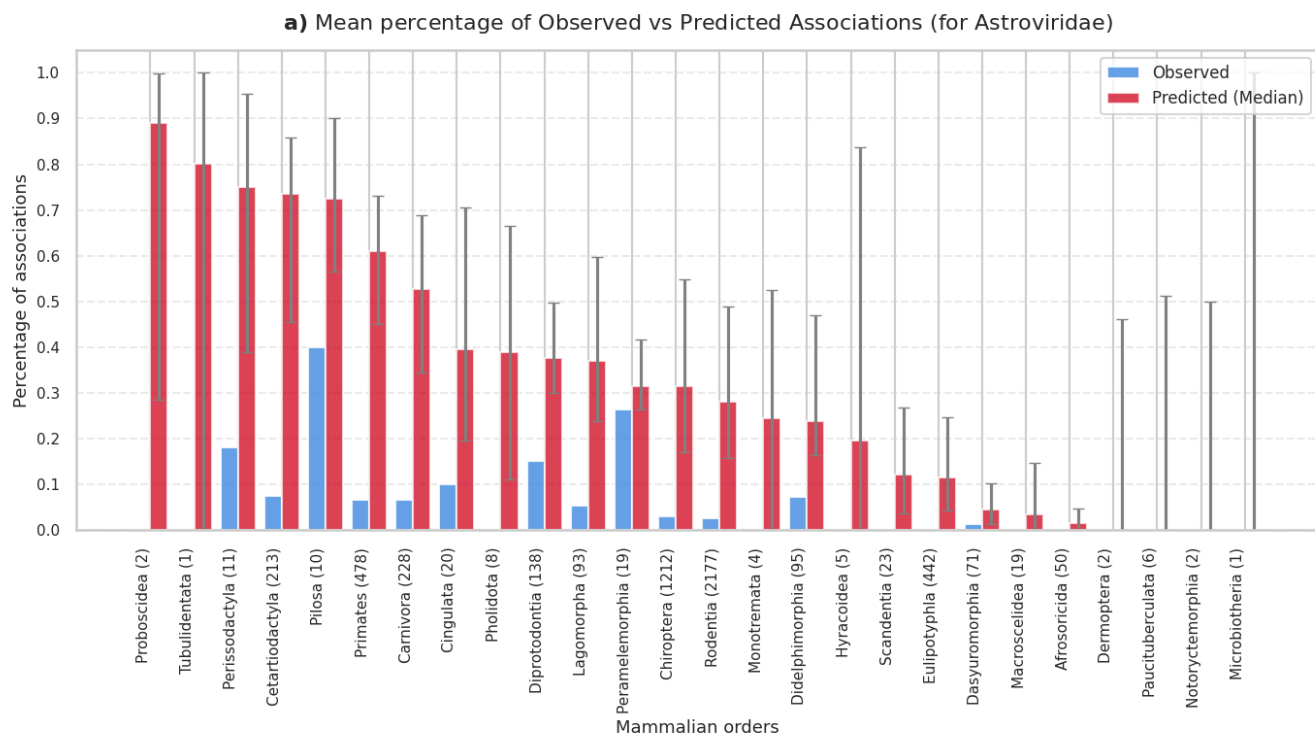
Supplementary Figure SR28 | Predicted associations by mammalian order for Anelloviridae. Percentage (a) and total number (b) of predicted associations across mammalian orders. Orange bars represent the medians of expected percentages across ensemble members, with uncertainty bars indicating the 5th to 95th percentile range. Blue bars indicate the observed percentages.



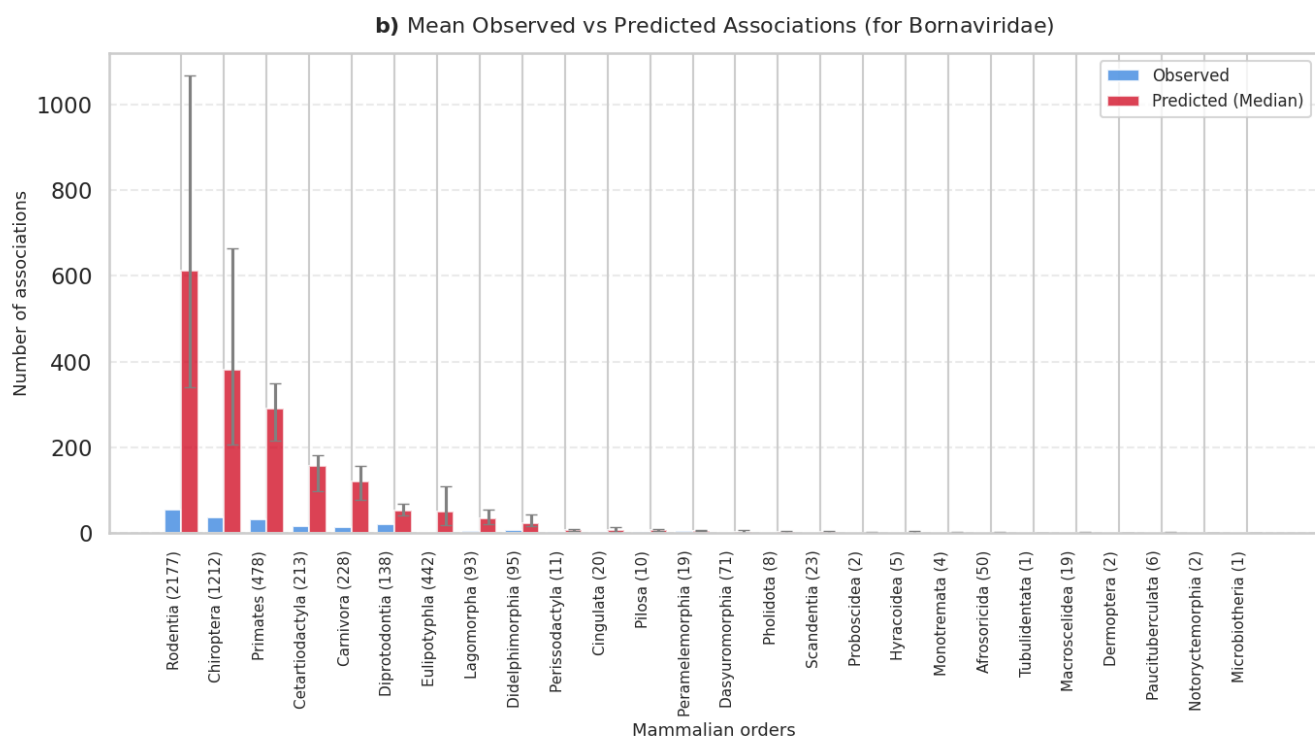
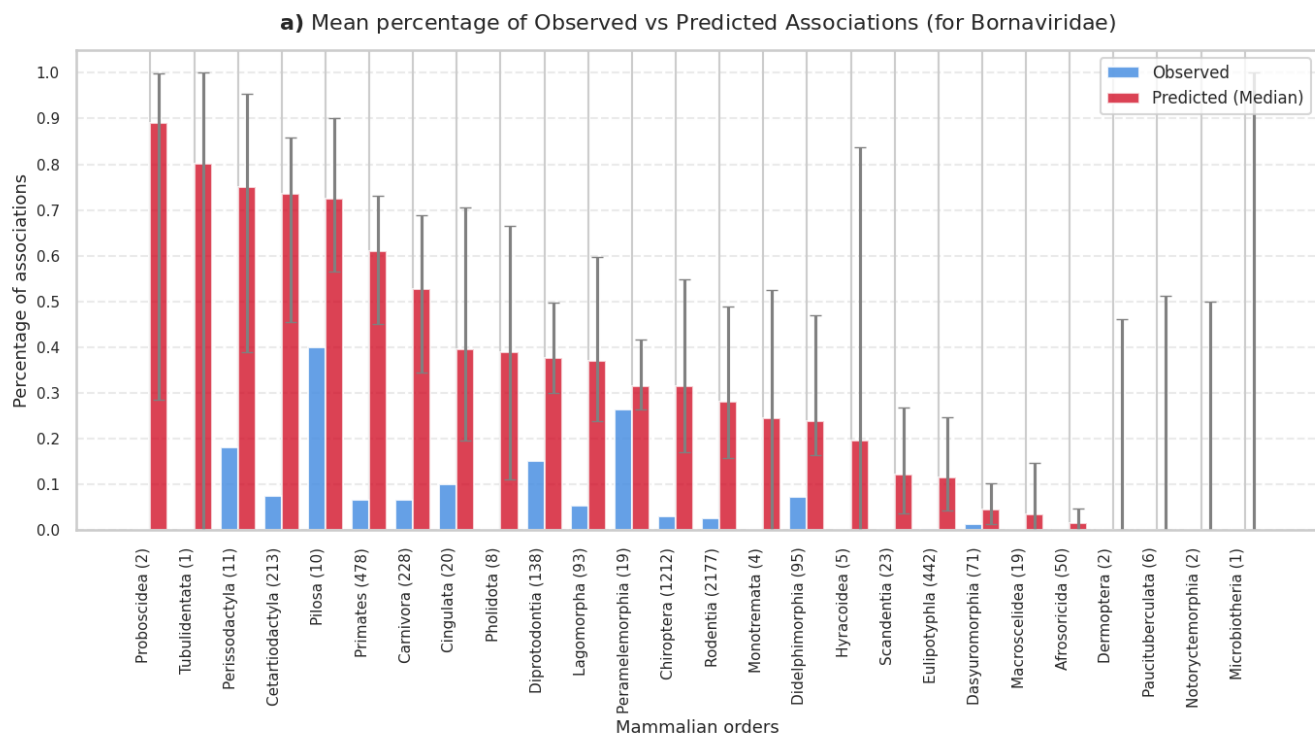
Supplementary Figure SR29 | Predicted associations by mammalian order for Arenaviridae. Percentage (a) and total number (b) of predicted associations across mammalian orders. Orange bars represent the medians of expected percentages across ensemble members, with uncertainty bars indicating the 5th to 95th percentile range. Blue bars indicate the observed percentages.



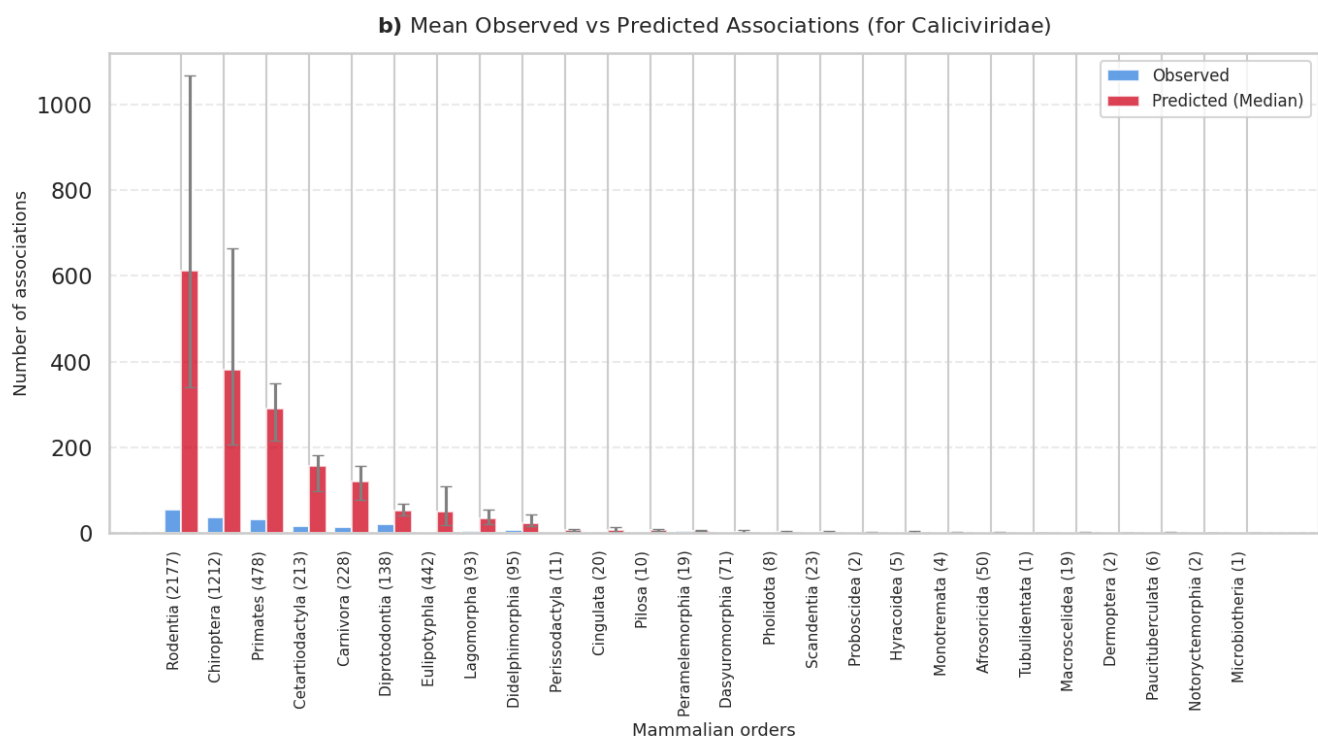
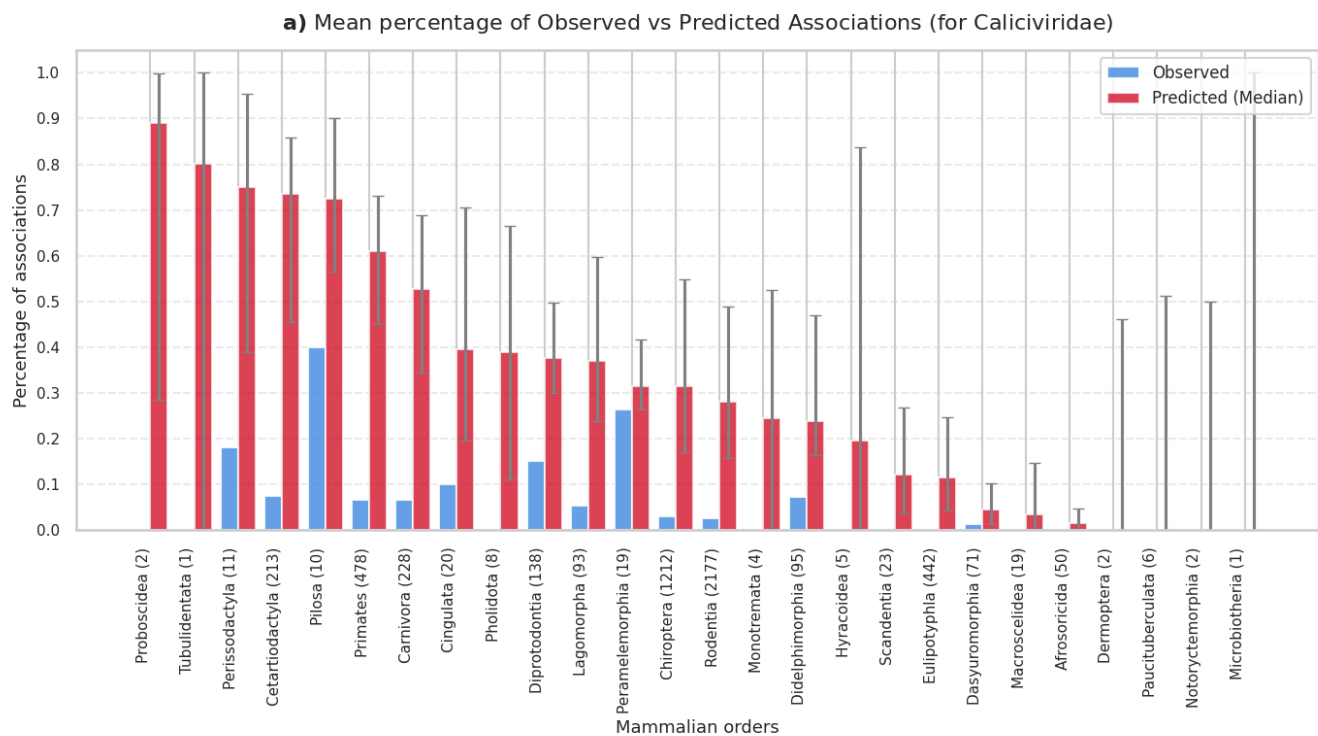
Supplementary Figure SR30 | Predicted associations by mammalian order for Arteriviridae. Percentage (a) and total number (b) of predicted associations across mammalian orders. Orange bars represent the medians of expected percentages across ensemble members, with uncertainty bars indicating the 5th to 95th percentile range. Blue bars indicate the observed percentages.



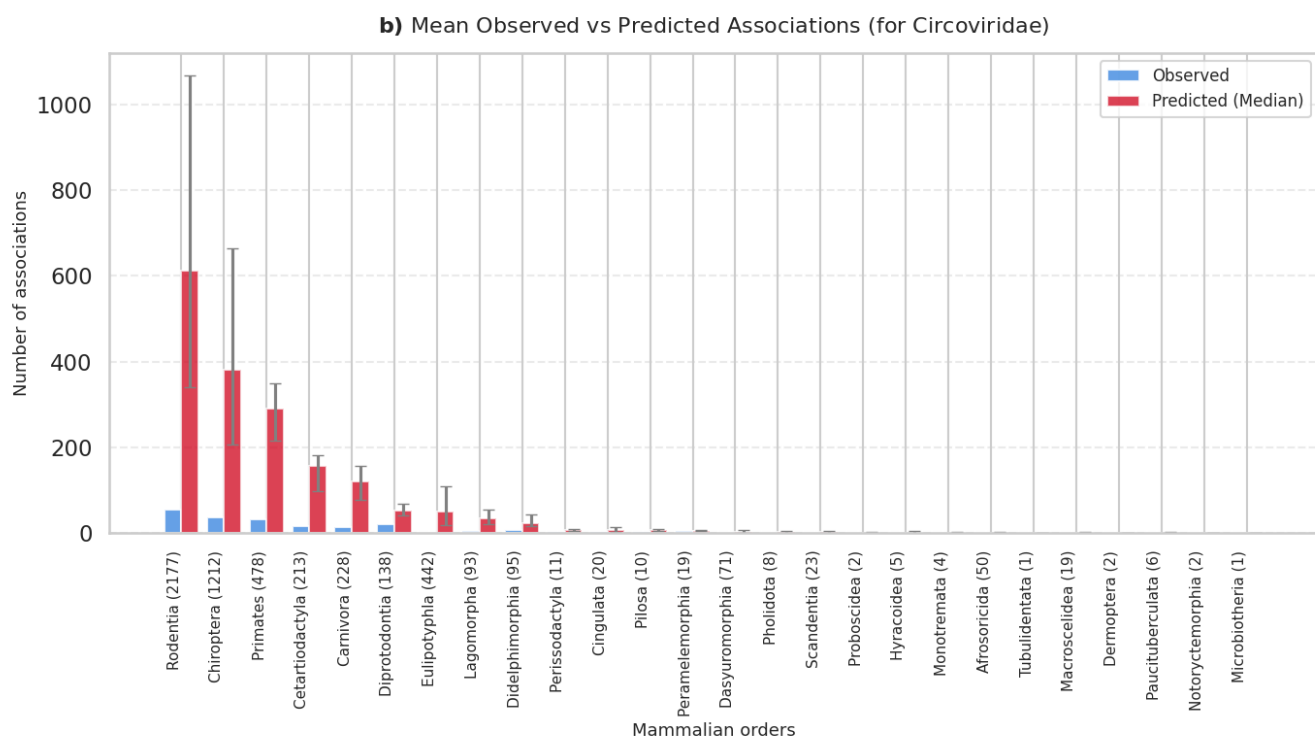
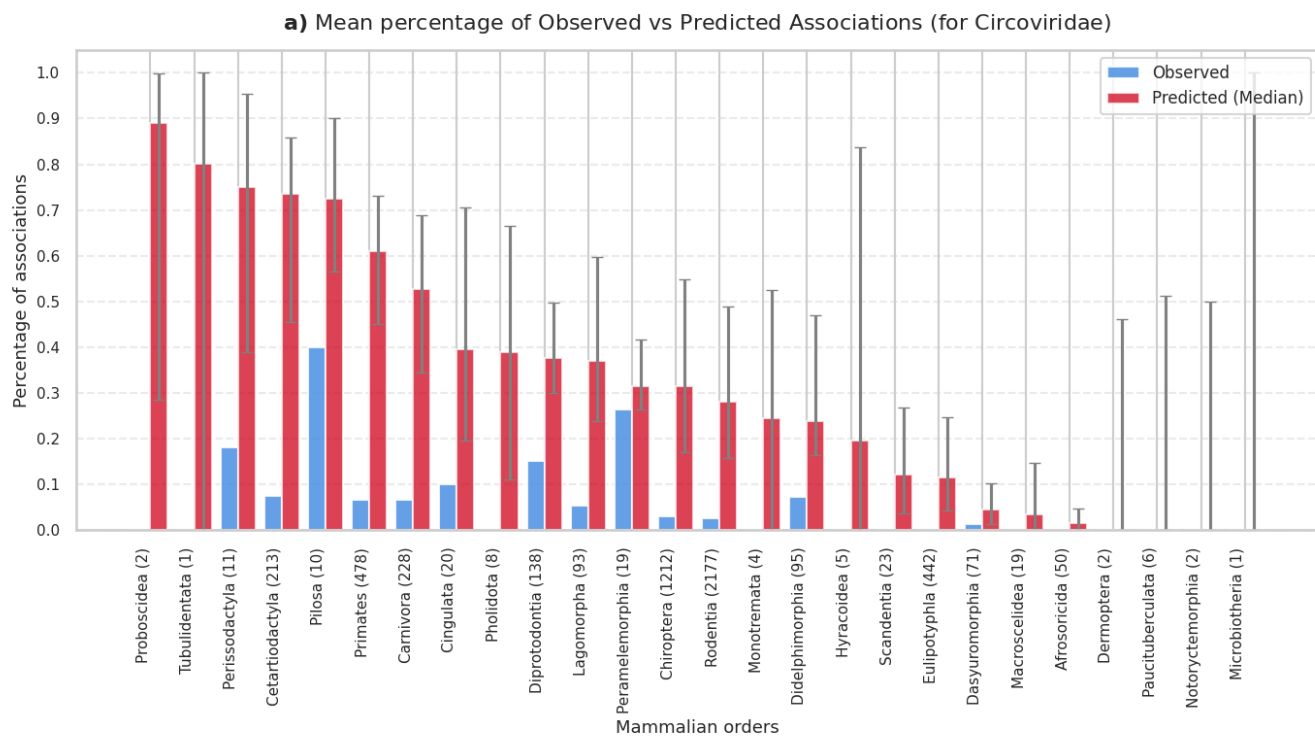
Supplementary Figure SR31 | Predicted associations by mammalian order for Astroviridae. Percentage (a) and total number (b) of predicted associations across mammalian orders. Orange bars represent the medians of expected percentages across ensemble members, with uncertainty bars indicating the 5th to 95th percentile range. Blue bars indicate the observed percentages.



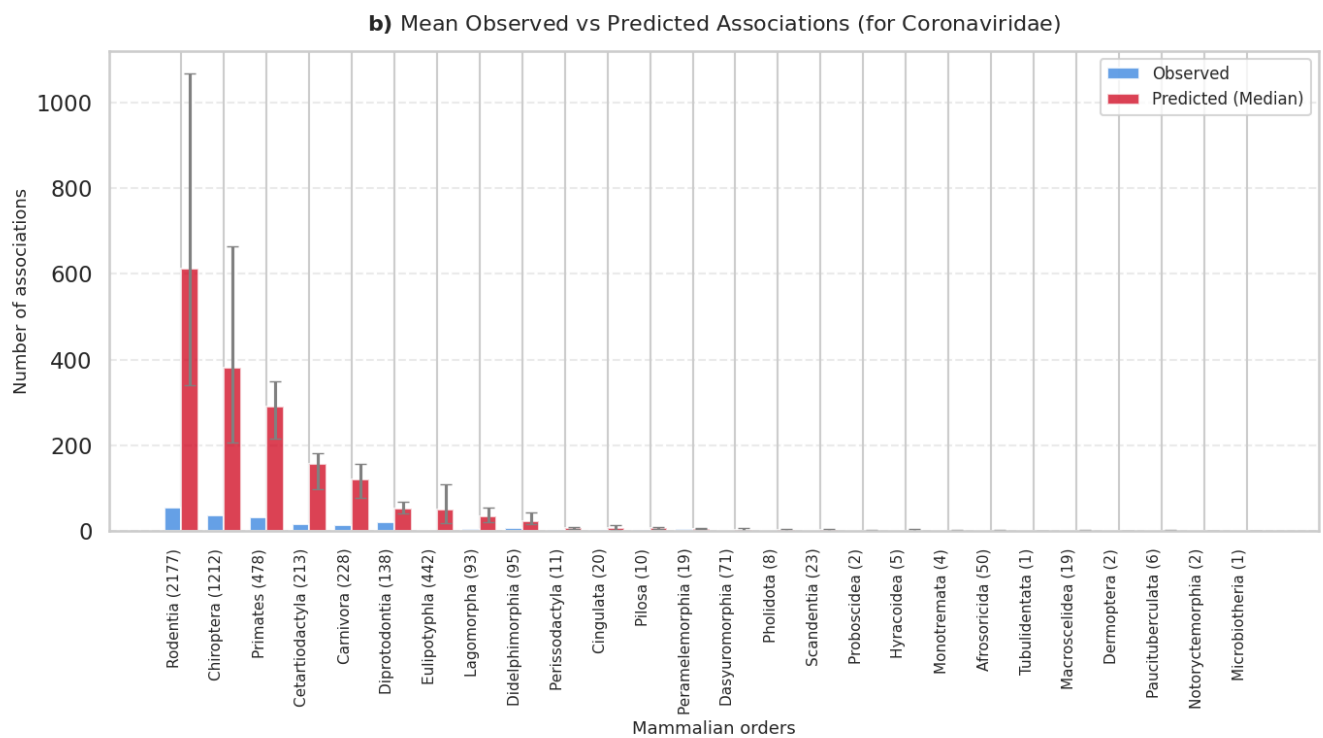
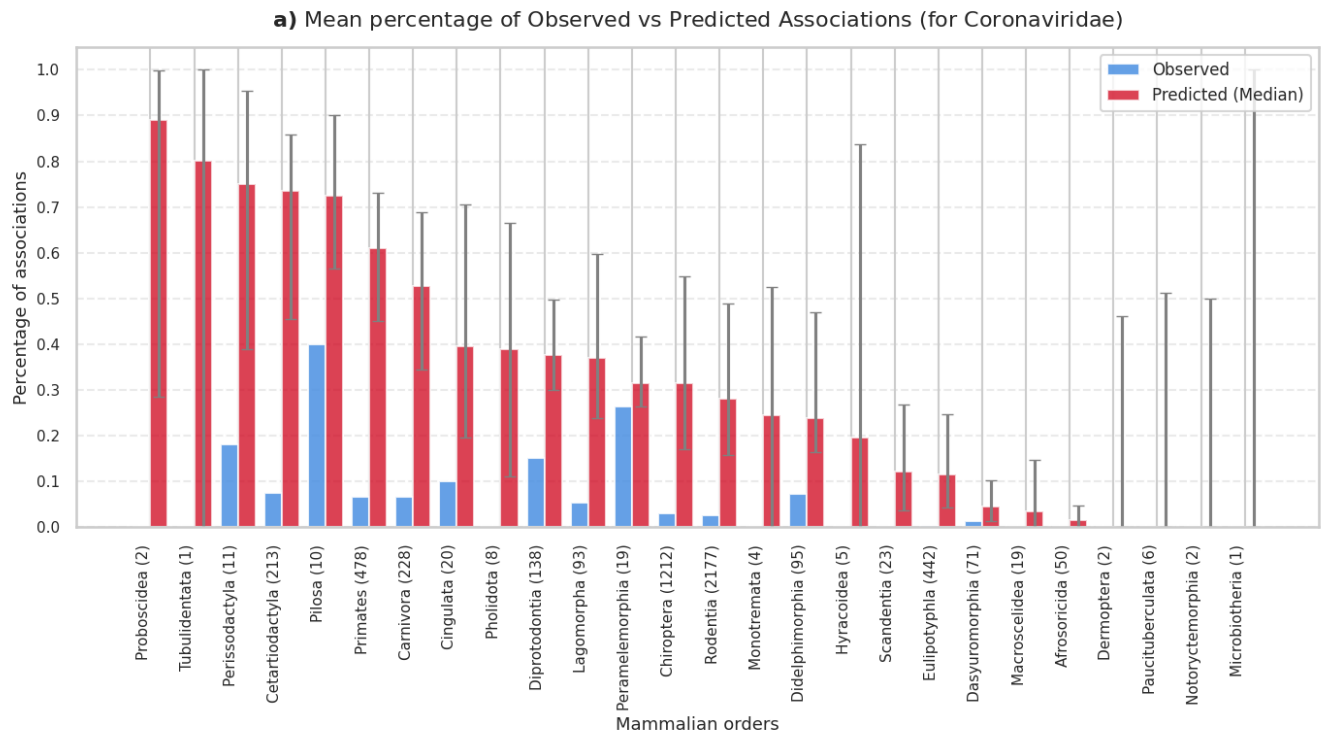
Supplementary Figure SR32 | Predicted associations by mammalian order for Bornaviridae. Percentage (a) and total number (b) of predicted associations across mammalian orders. Orange bars represent the medians of expected percentages across ensemble members, with uncertainty bars indicating the 5th to 95th percentile range. Blue bars indicate the observed percentages.



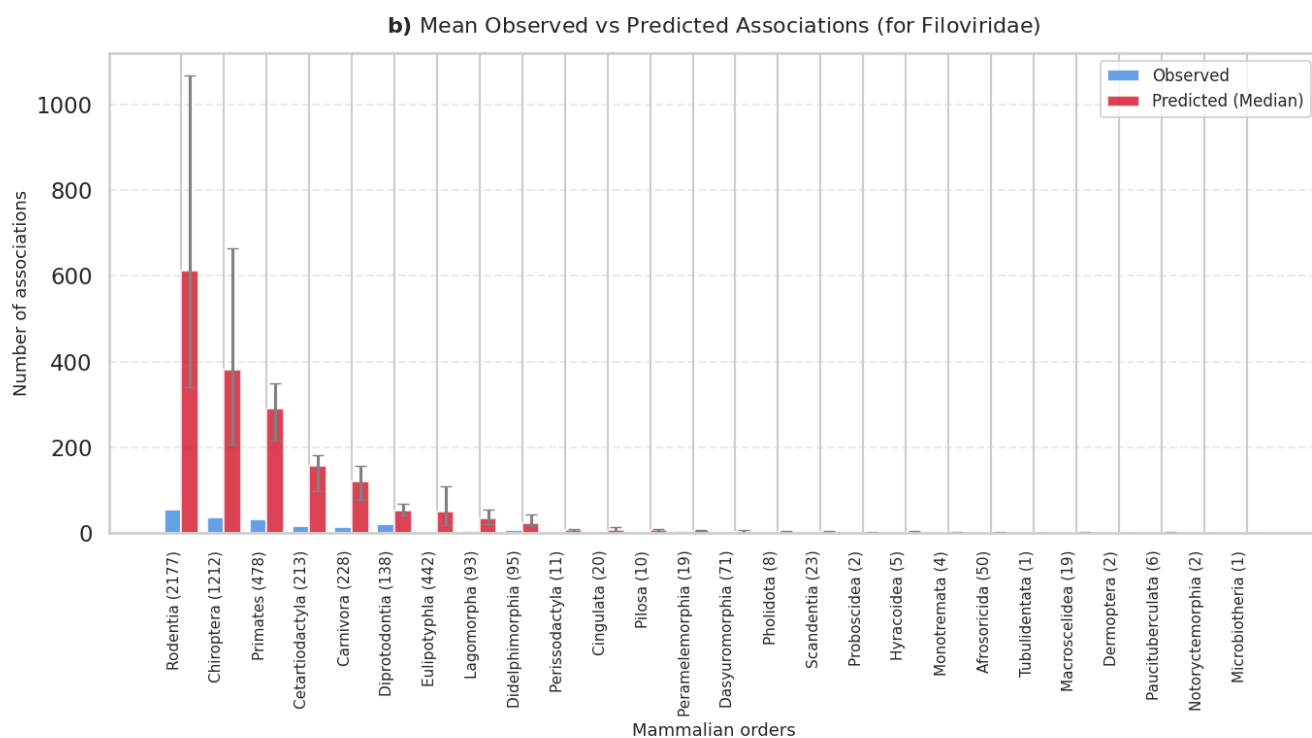
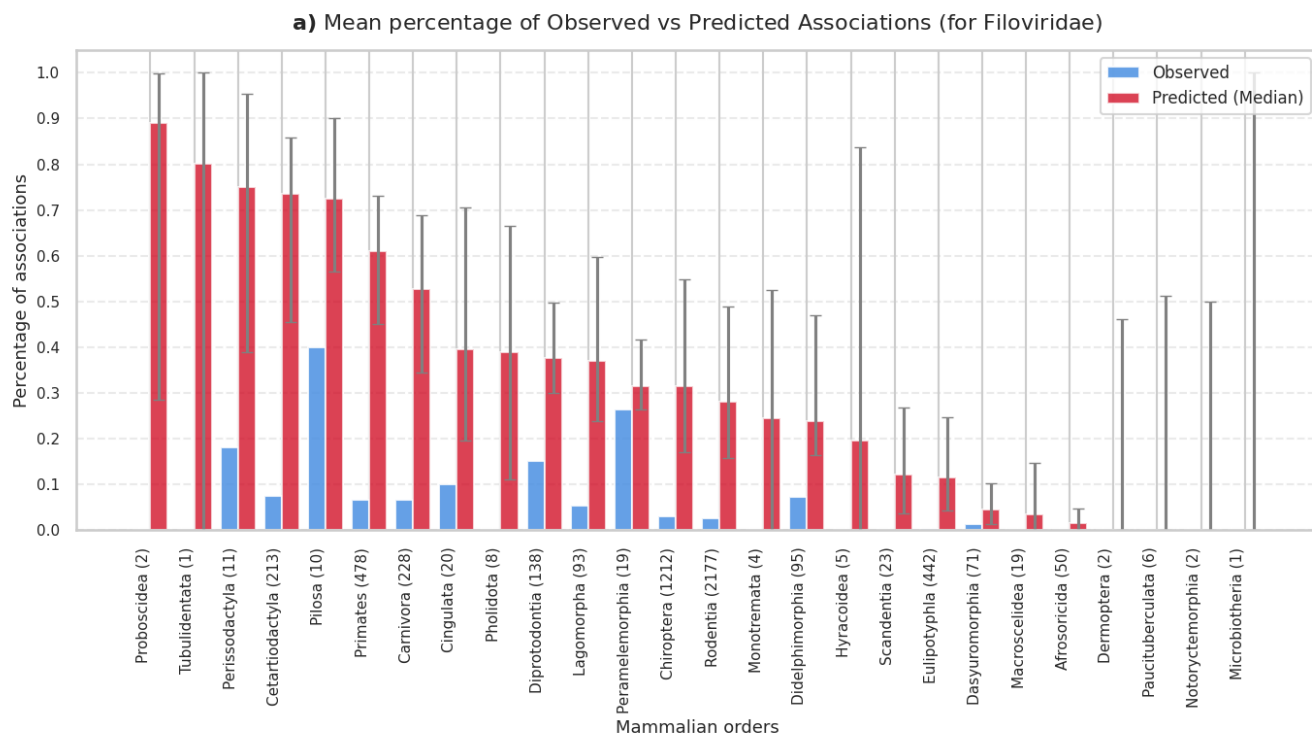
Supplementary Figure SR33 | Predicted associations by mammalian order for Caliciviridae. Percentage (a) and total number (b) of predicted associations across mammalian orders. Orange bars represent the medians of expected percentages across ensemble members, with uncertainty bars indicating the 5th to 95th percentile range. Blue bars indicate the observed percentages.



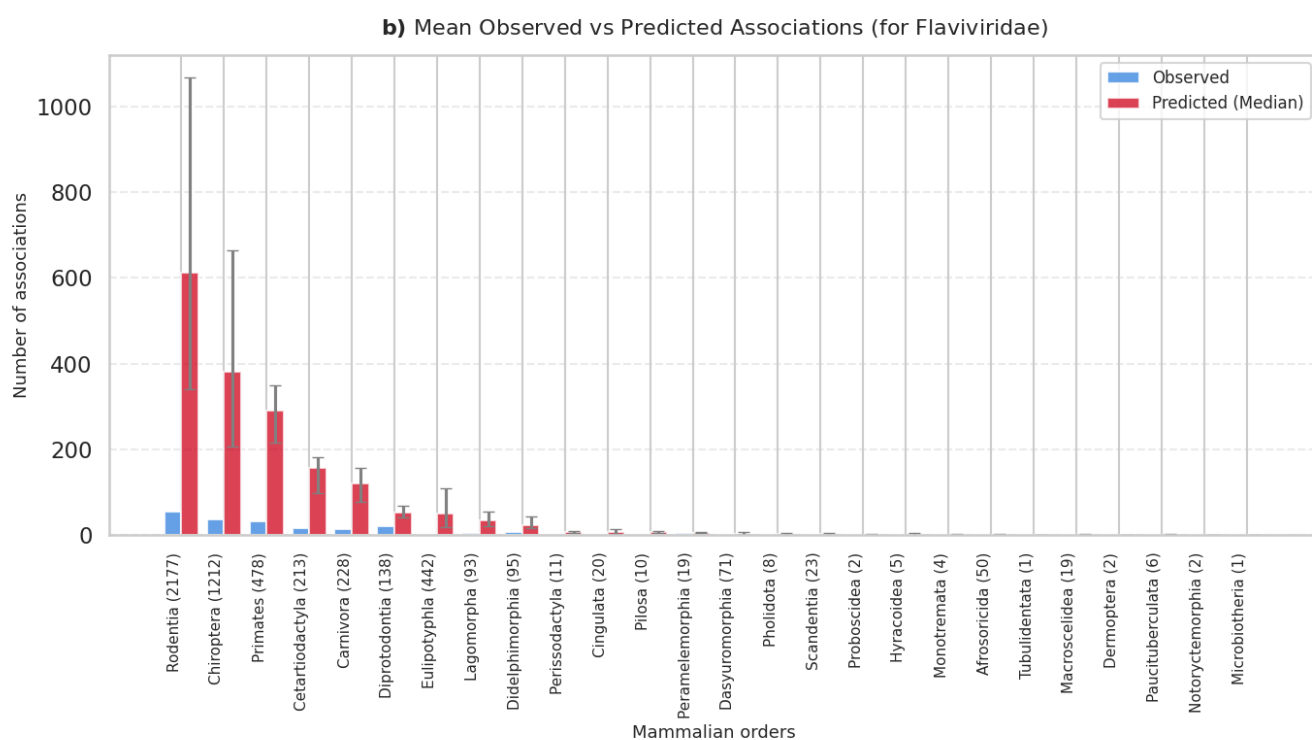
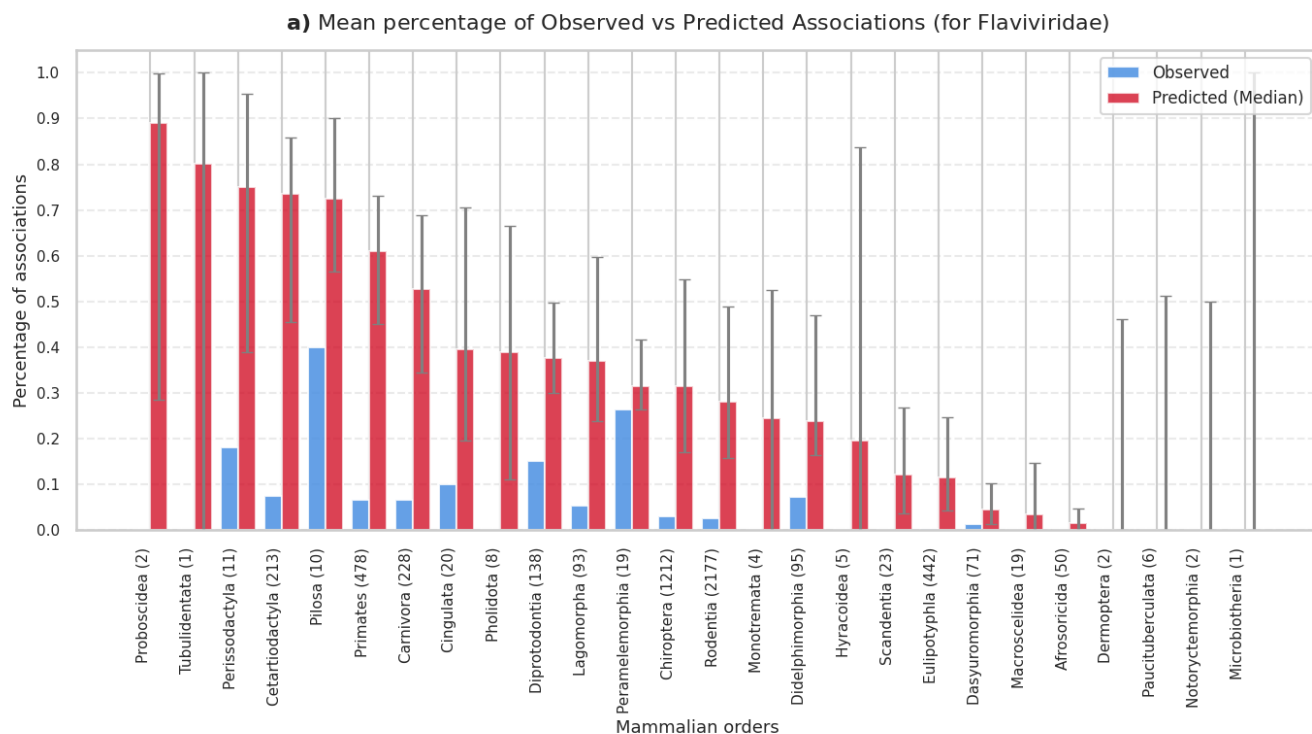
Supplementary Figure SR34 | Predicted associations by mammalian order for Circoviridae. Percentage (a) and total number (b) of predicted associations across mammalian orders. Orange bars represent the medians of expected percentages across ensemble members, with uncertainty bars indicating the 5th to 95th percentile range. Blue bars indicate the observed percentages.



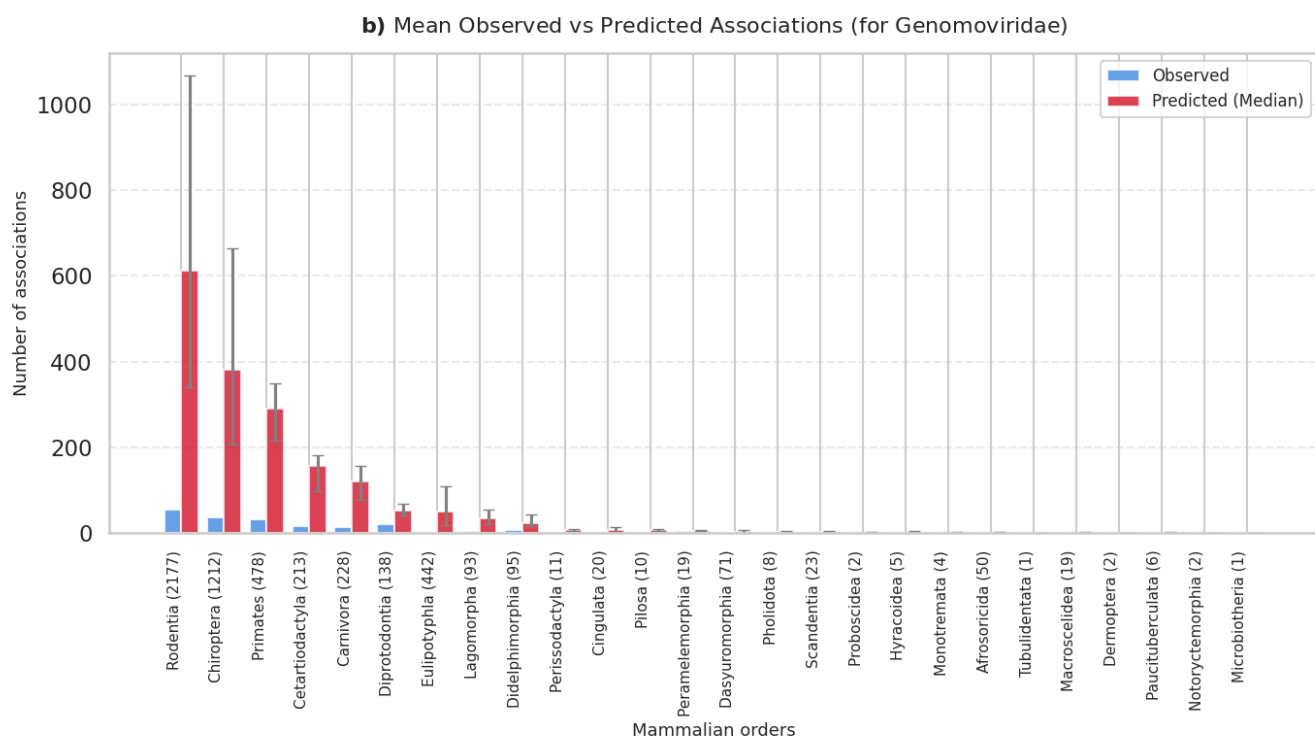
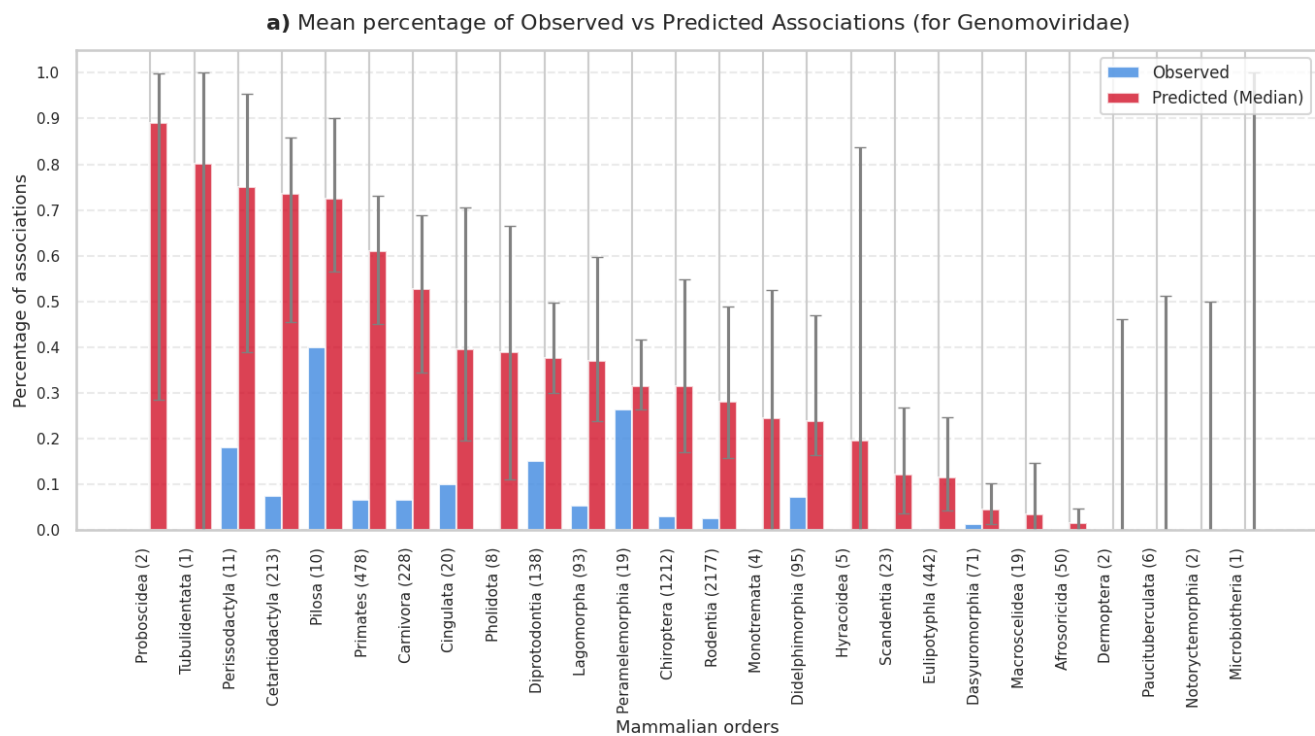
Supplementary Figure SR35 | Predicted associations by mammalian order for Coronaviridae. Percentage (a) and total number (b) of predicted associations across mammalian orders. Orange bars represent the medians of expected percentages across ensemble members, with uncertainty bars indicating the 5th to 95th percentile range. Blue bars indicate the observed percentages.



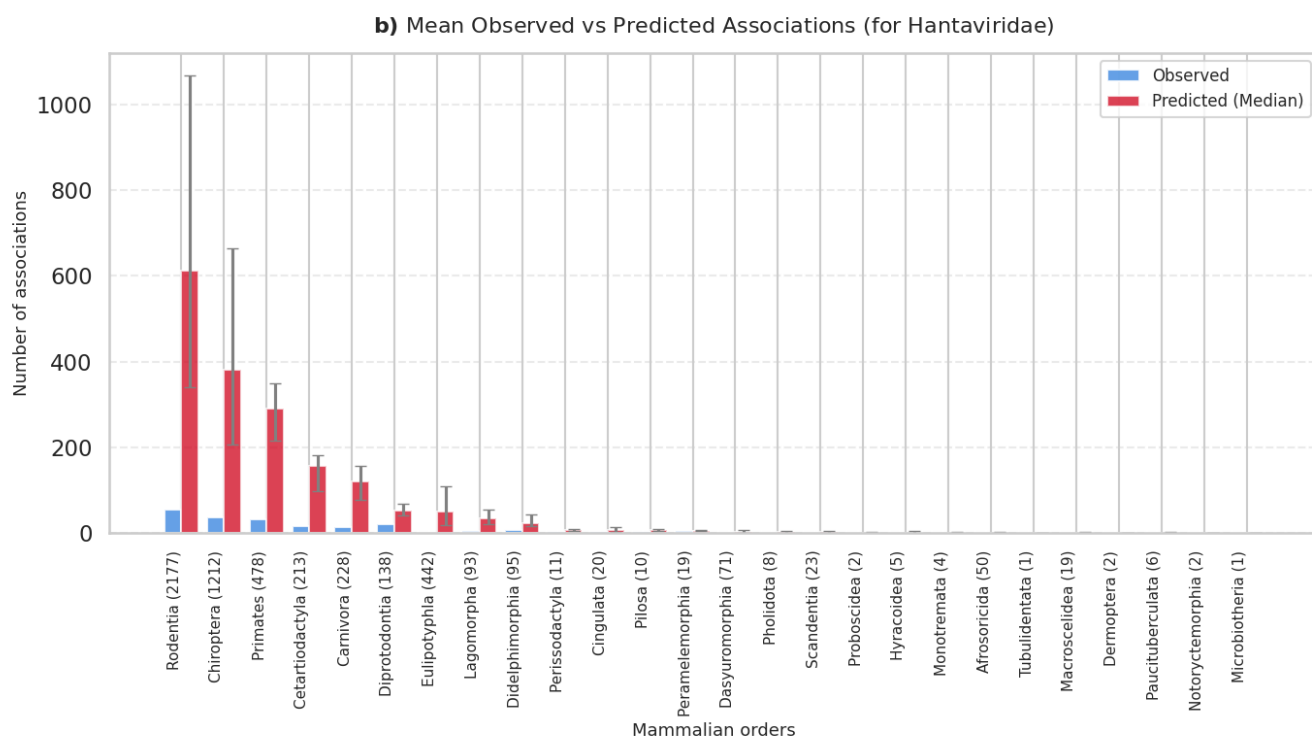
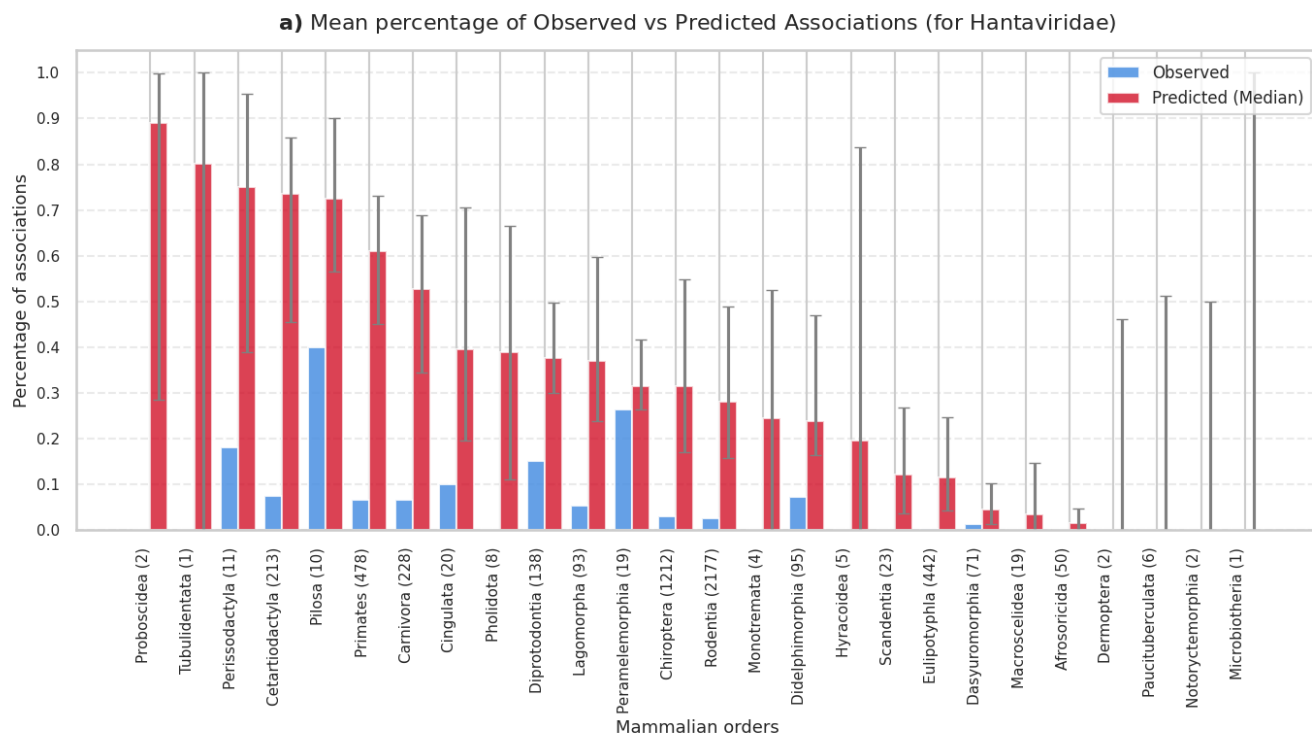
Supplementary Figure SR36 | Predicted associations by mammalian order for Filoviridae. Percentage (a) and total number (b) of predicted associations across mammalian orders. Orange bars represent the medians of expected percentages across ensemble members, with uncertainty bars indicating the 5th to 95th percentile range. Blue bars indicate the observed percentages.



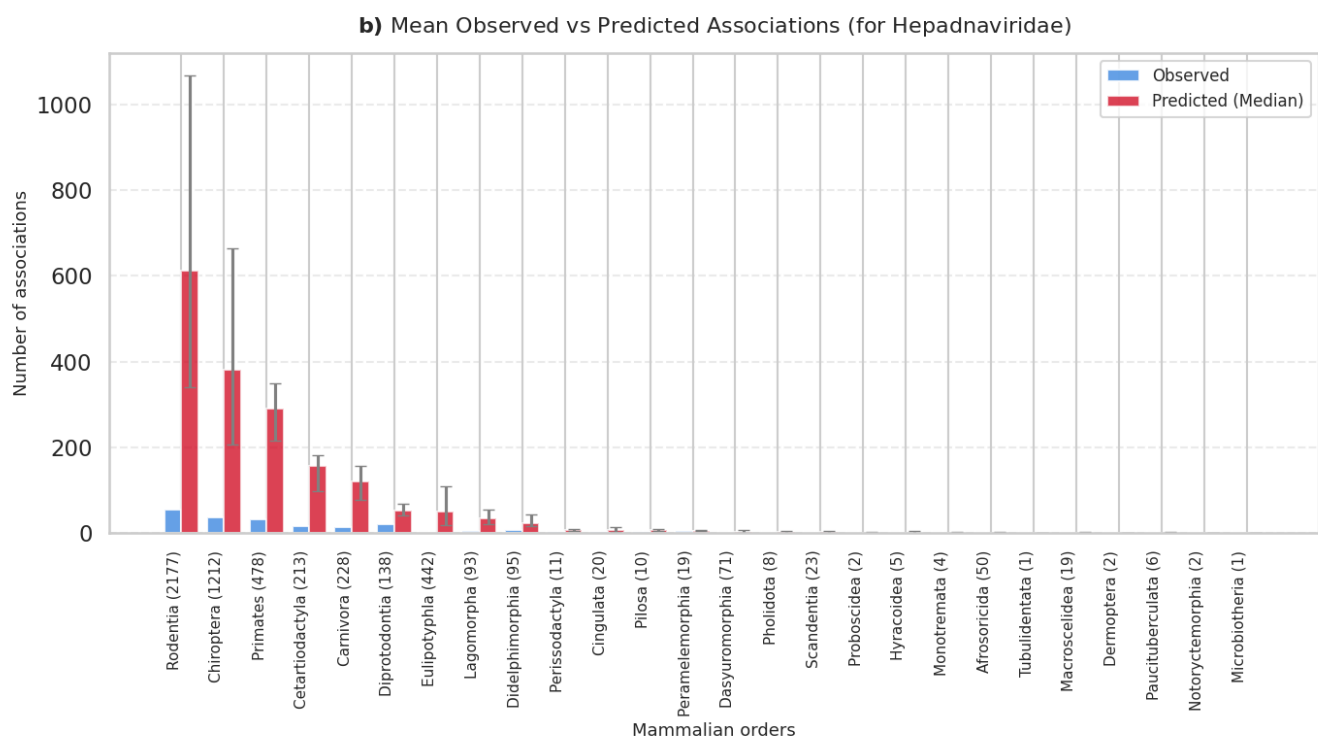
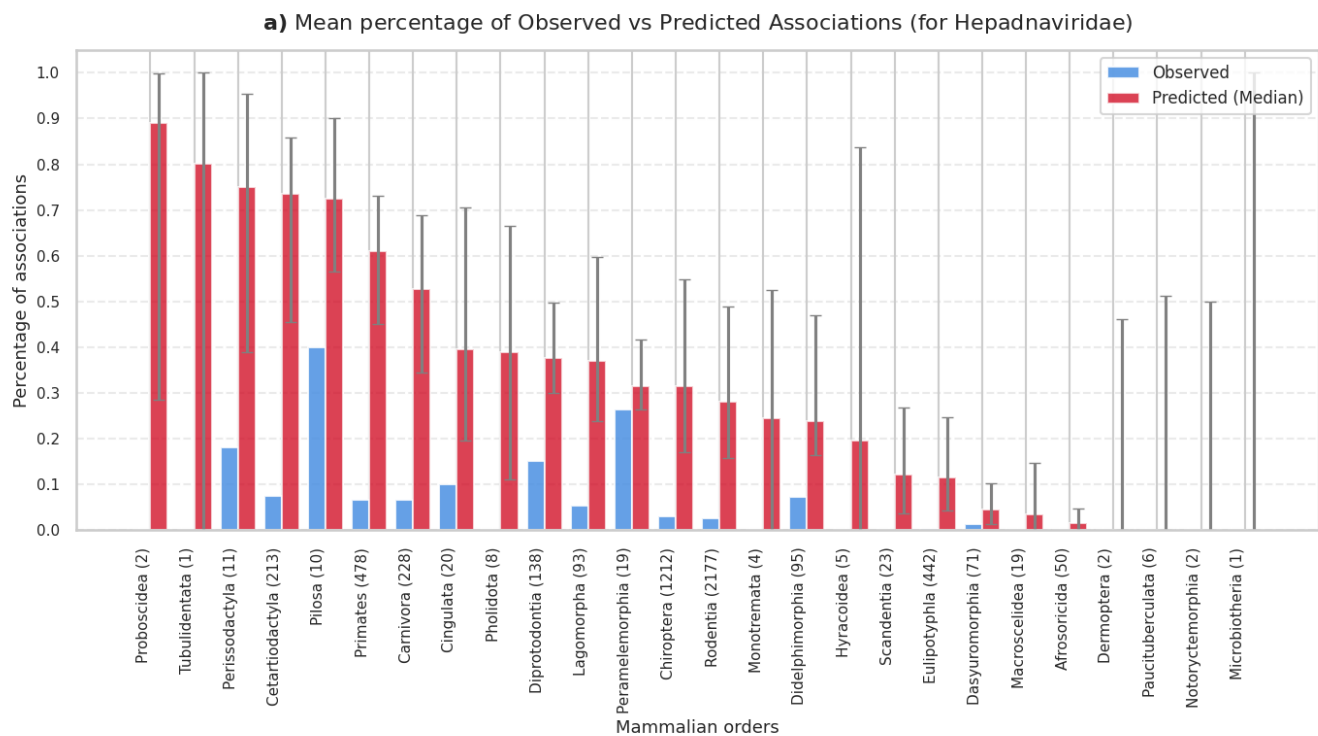
Supplementary Figure SR37 | Predicted associations by mammalian order for Flaviviridae. Percentage (a) and total number (b) of predicted associations across mammalian orders. Orange bars represent the medians of expected percentages across ensemble members, with uncertainty bars indicating the 5th to 95th percentile range. Blue bars indicate the observed percentages.



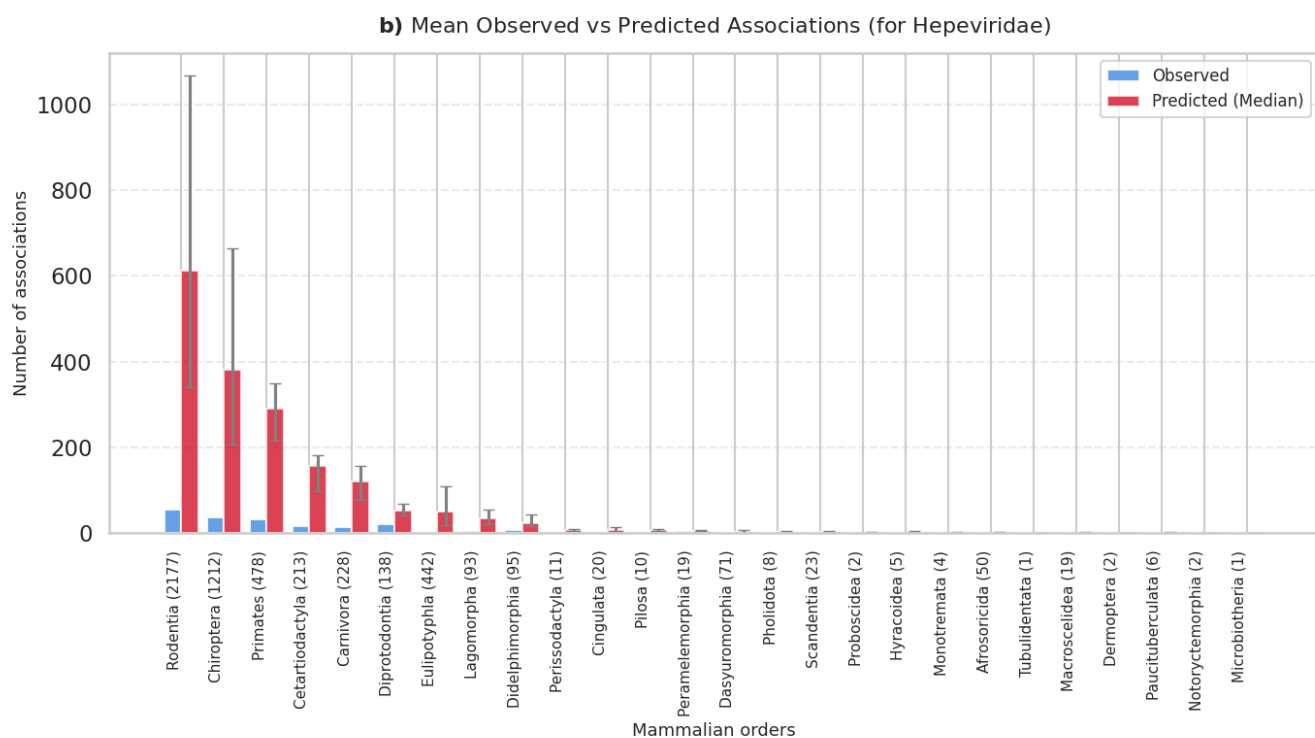
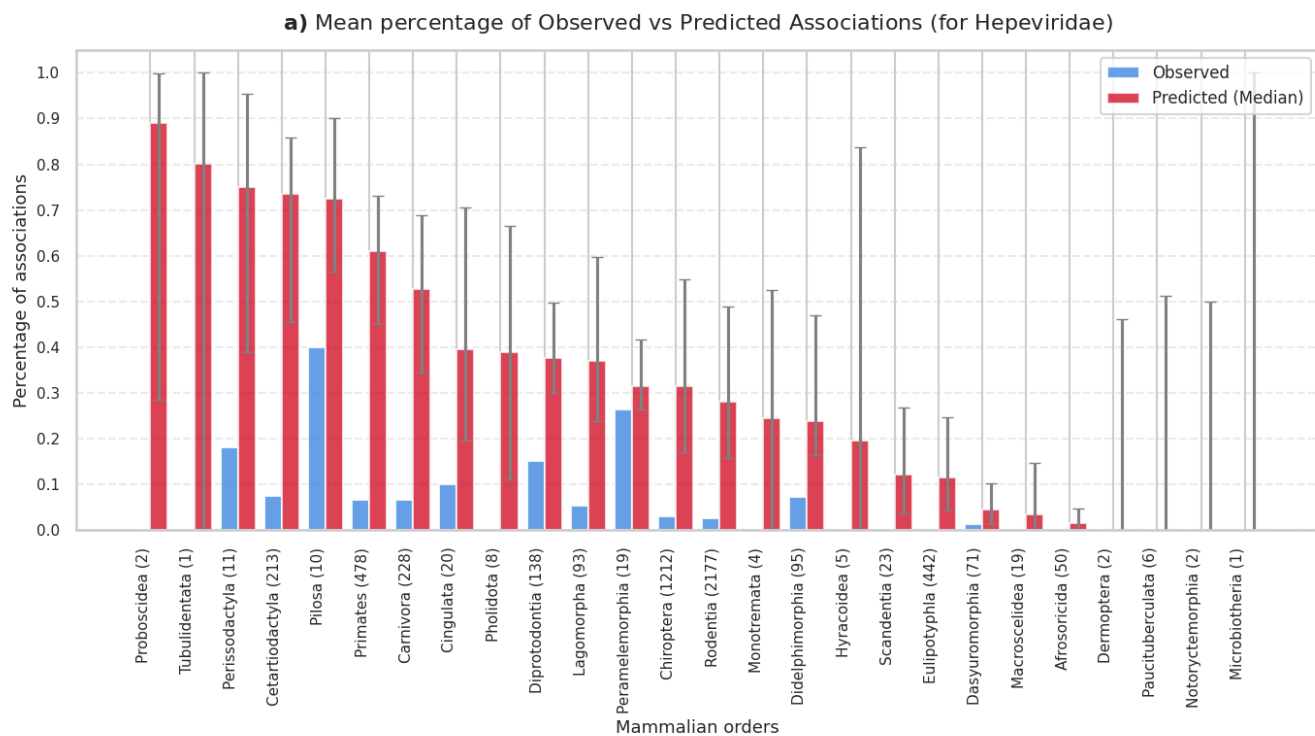
Supplementary Figure SR38 | Predicted associations by mammalian order for Genomoviridae. Percentage (a) and total number (b) of predicted associations across mammalian orders. Orange bars represent the medians of expected percentages across ensemble members, with uncertainty bars indicating the 5th to 95th percentile range. Blue bars indicate the observed percentages.



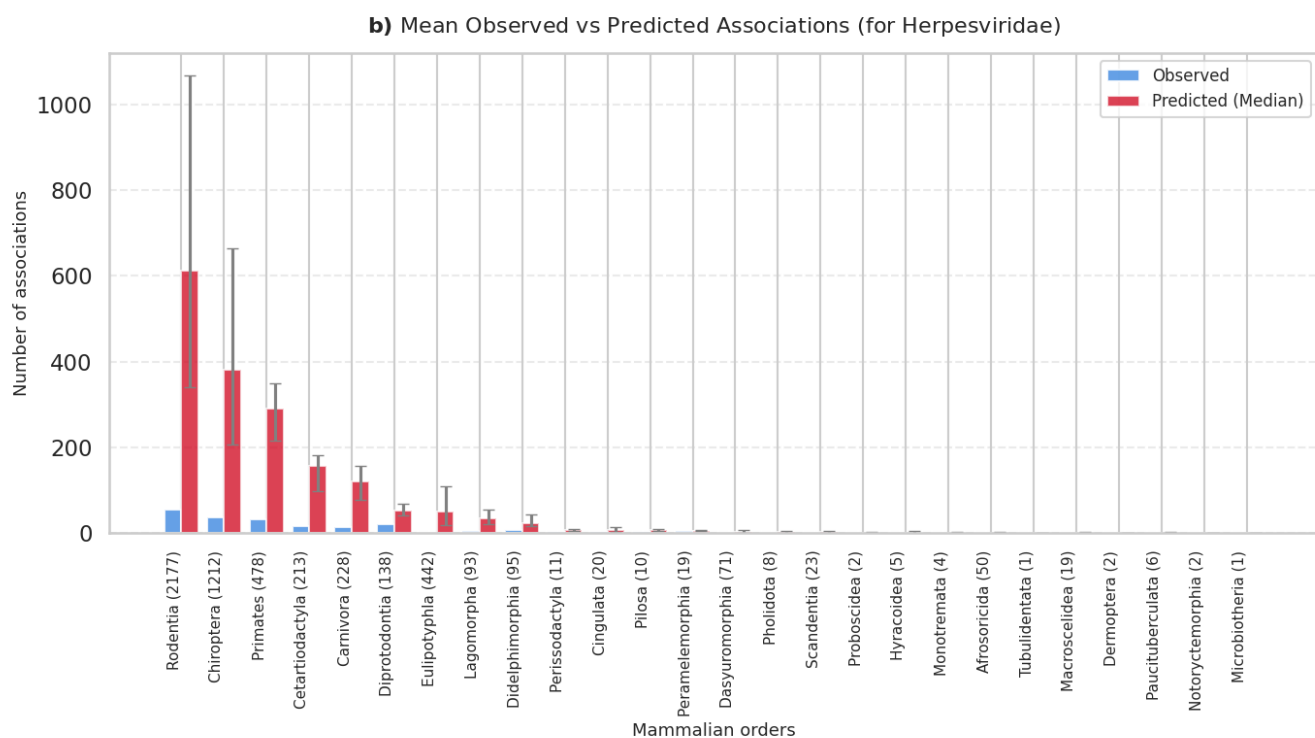
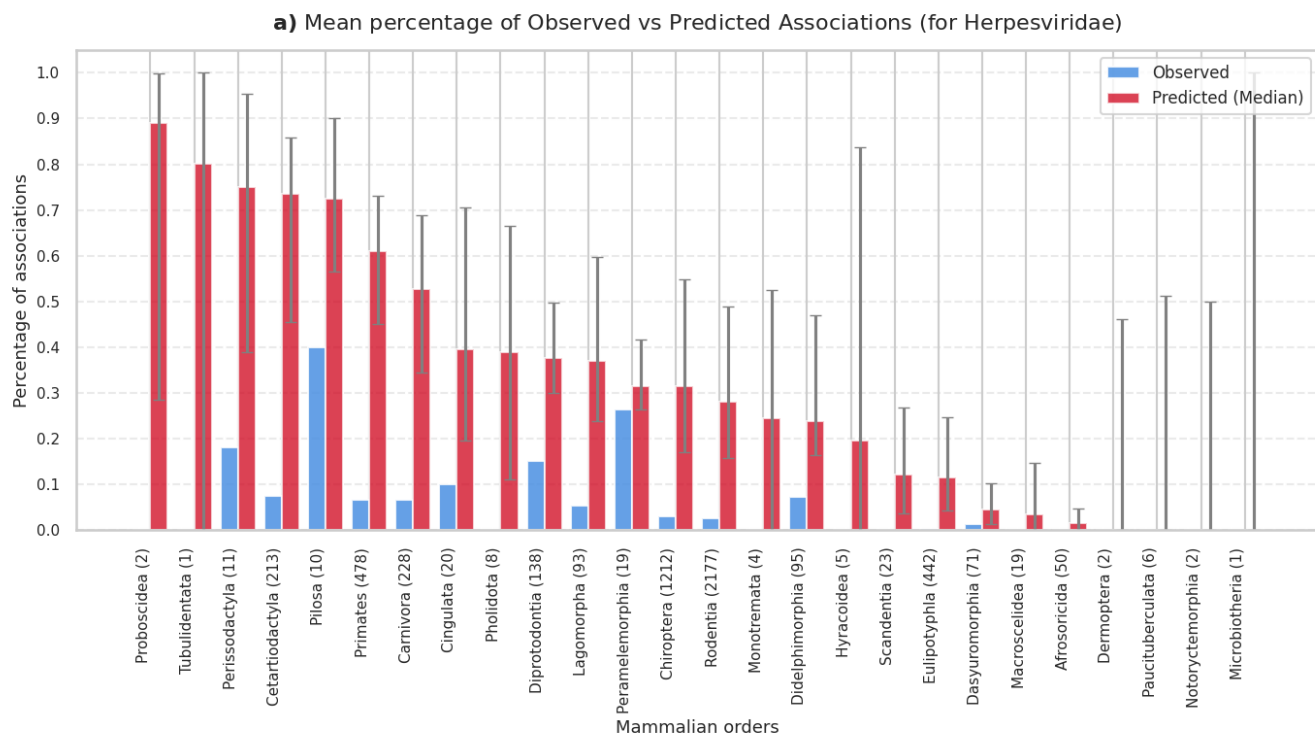
Supplementary Figure SR39 | Predicted associations by mammalian order for Hantaviridae. Percentage (a) and total number (b) of predicted associations across mammalian orders. Orange bars represent the medians of expected percentages across ensemble members, with uncertainty bars indicating the 5th to 95th percentile range. Blue bars indicate the observed percentages.



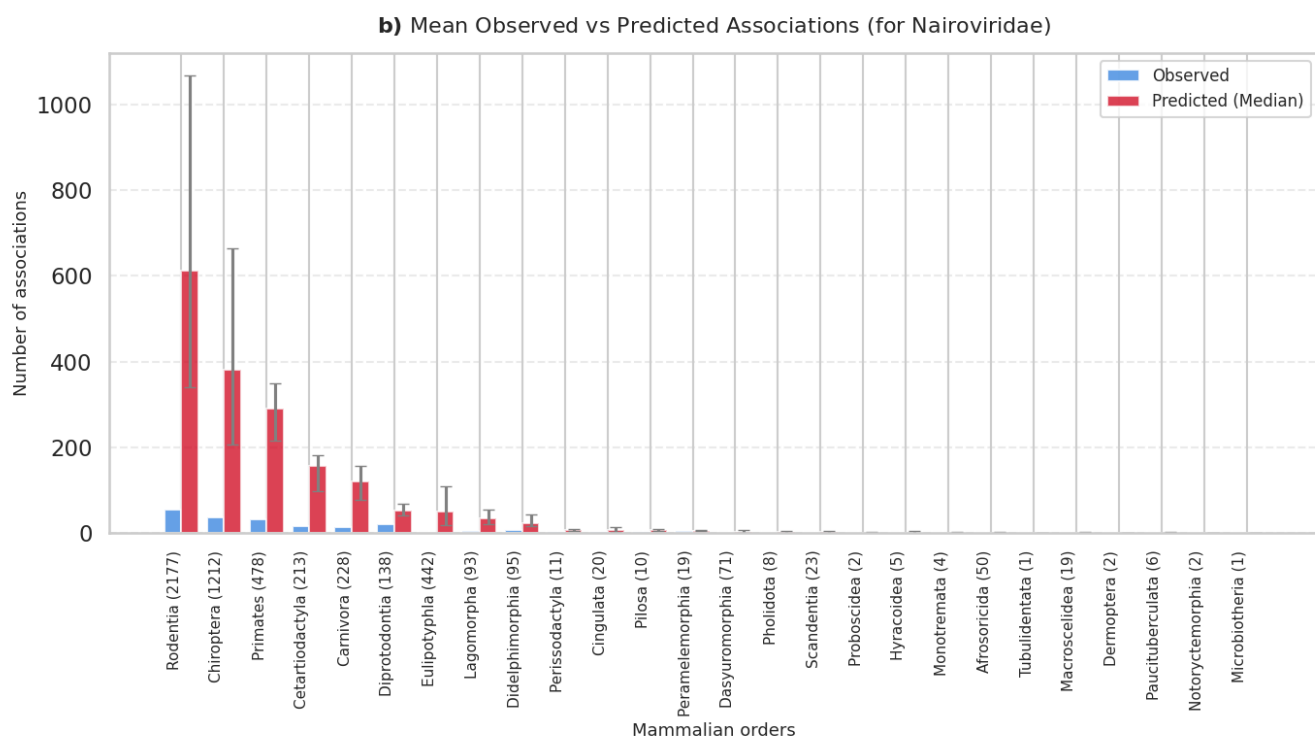
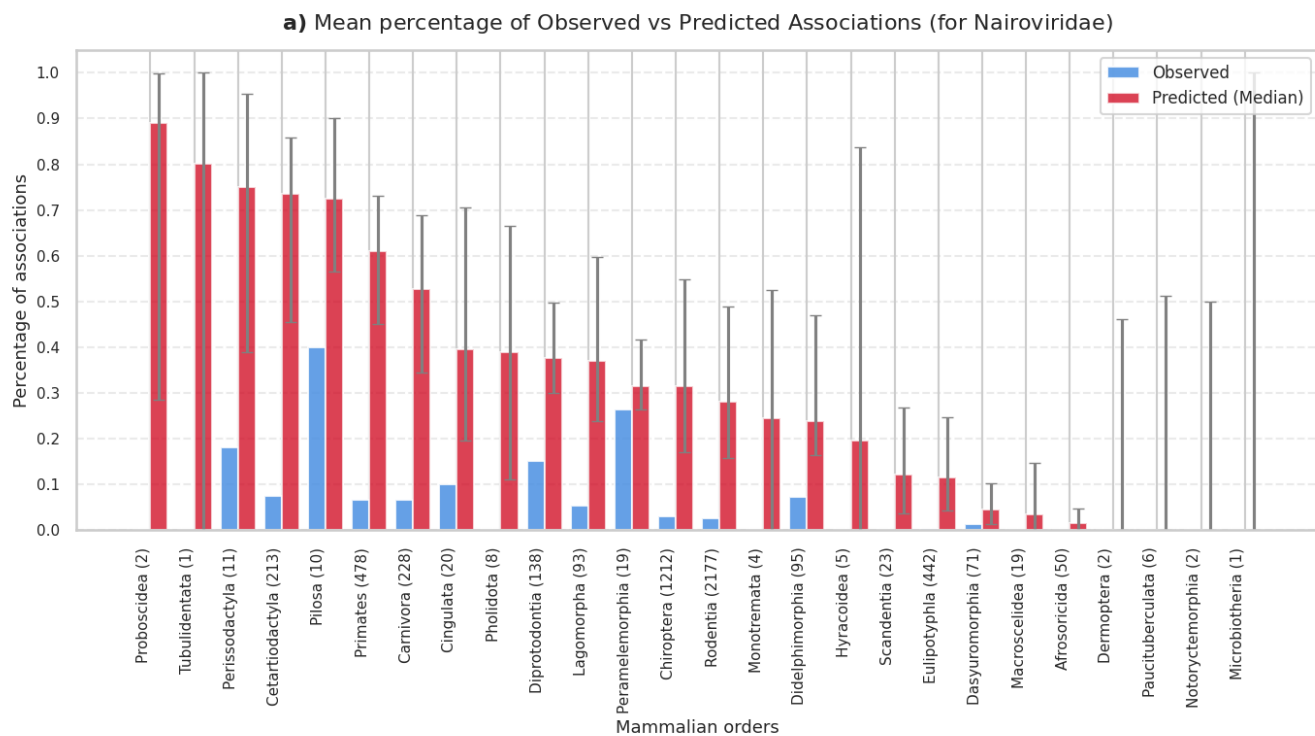
Supplementary Figure SR40 | Predicted associations by mammalian order for Hepadnaviridae. Percentage (a) and total number (b) of predicted associations across mammalian orders. Orange bars represent the medians of expected percentages across ensemble members, with uncertainty bars indicating the 5th to 95th percentile range. Blue bars indicate the observed percentages.



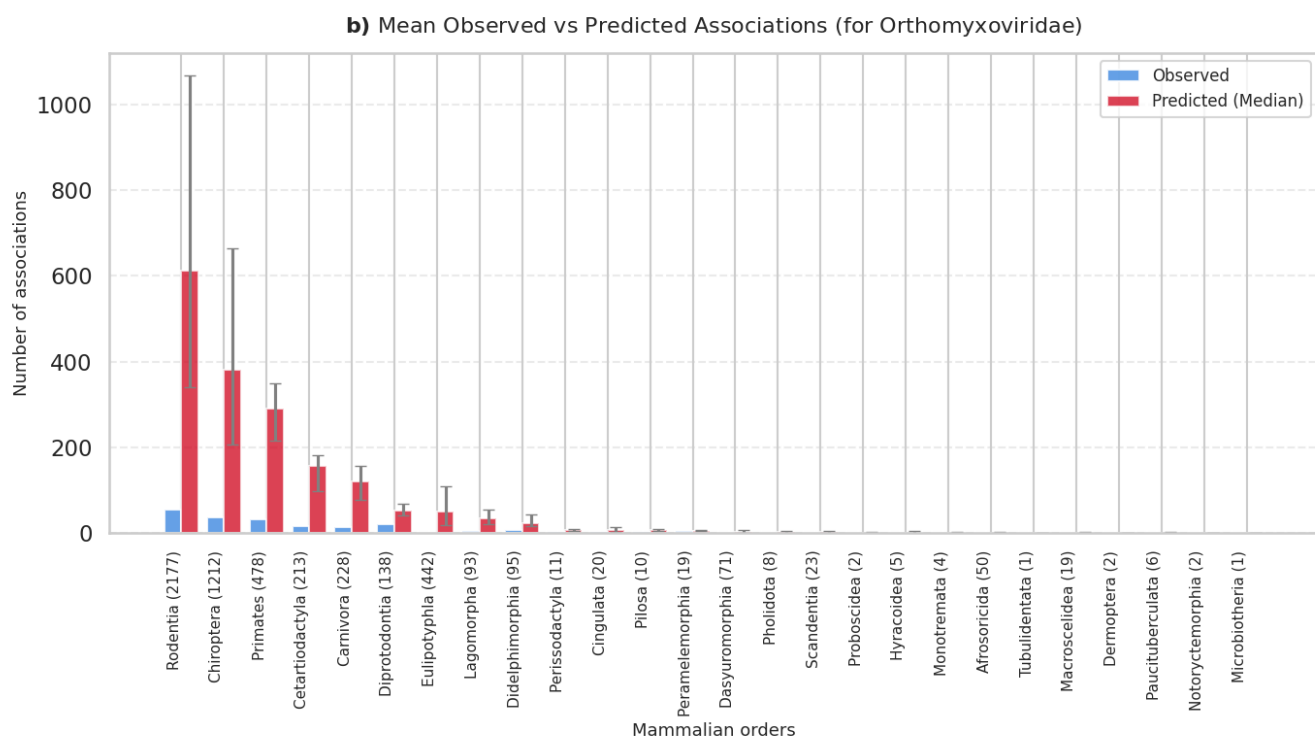
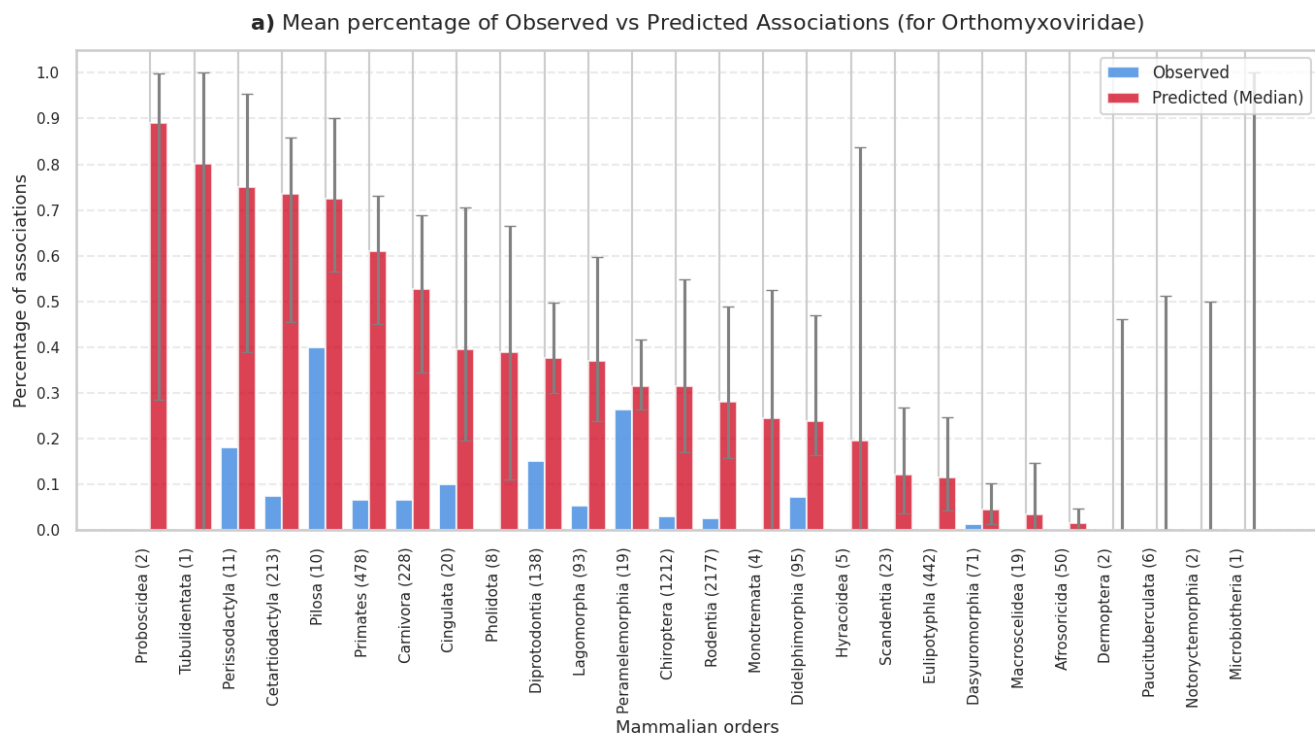
Supplementary Figure SR41 | Predicted associations by mammalian order for Hepeviridae. Percentage (a) and total number (b) of predicted associations across mammalian orders. Orange bars represent the medians of expected percentages across ensemble members, with uncertainty bars indicating the 5th to 95th percentile range. Blue bars indicate the observed percentages.



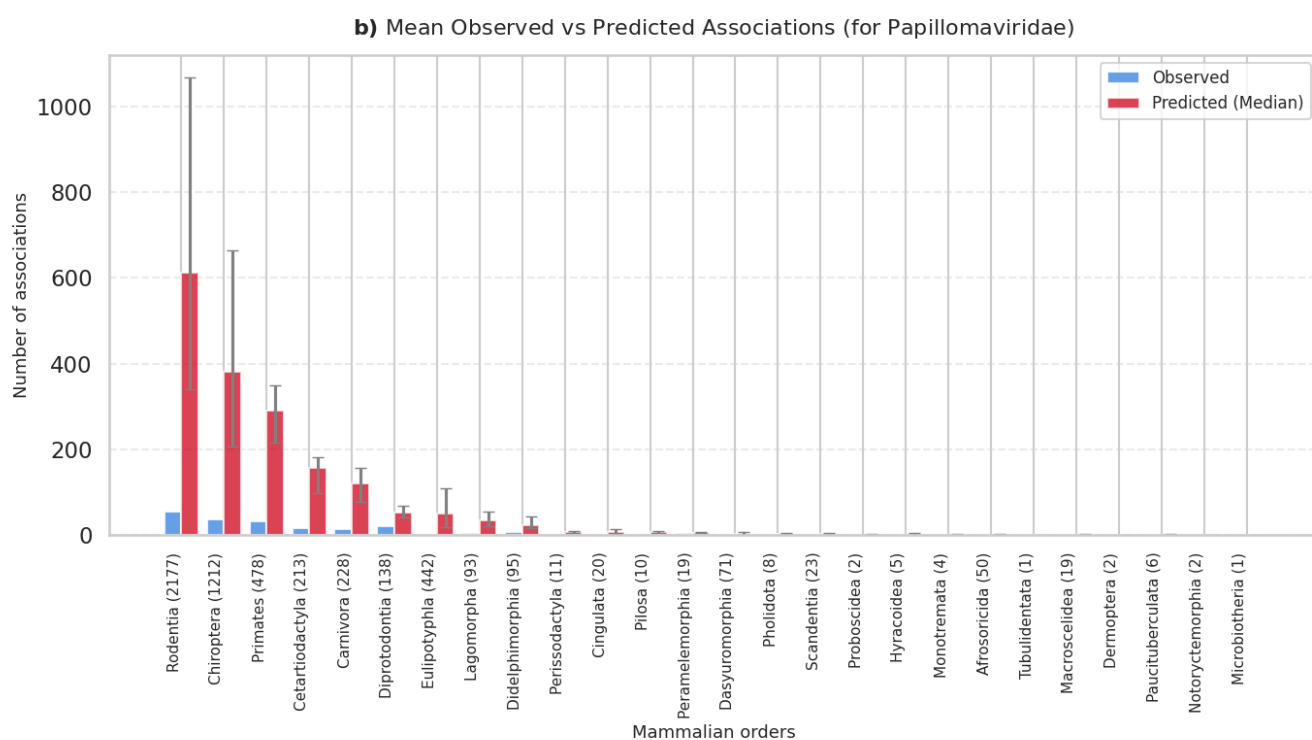
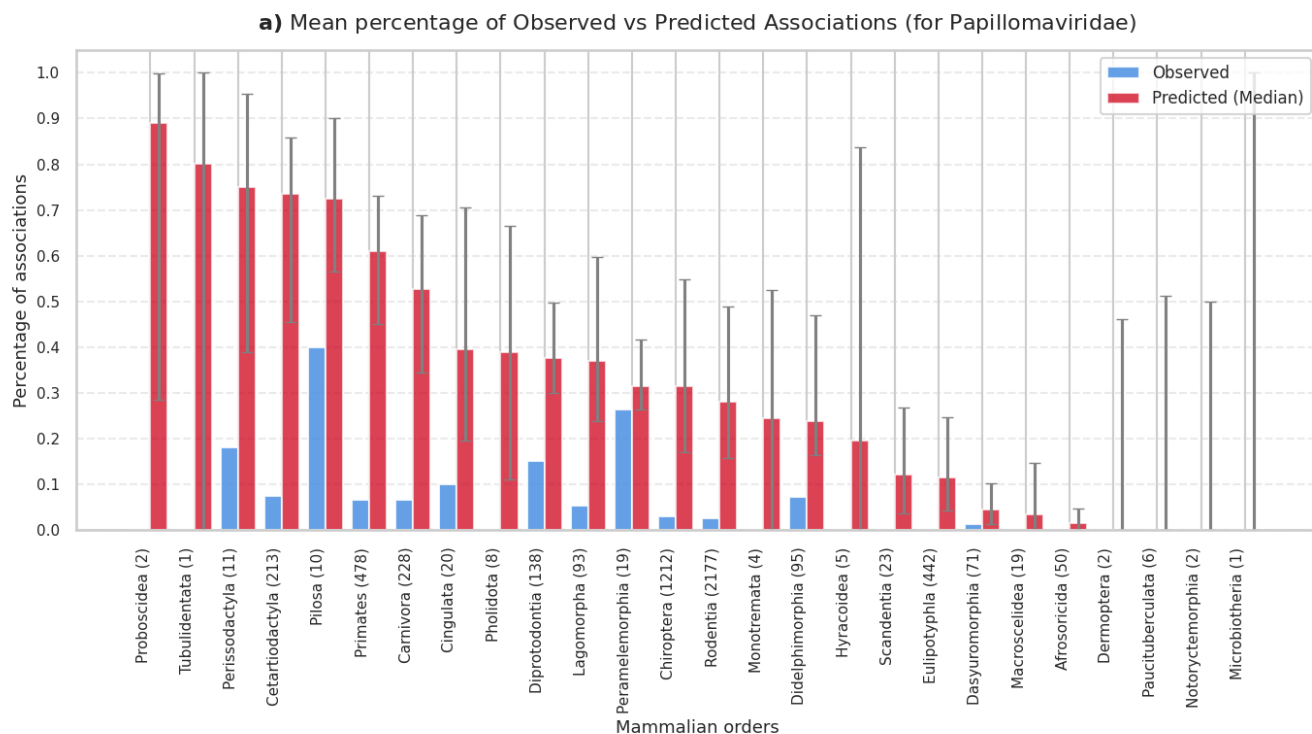
Supplementary Figure SR42 | Predicted associations by mammalian order for Herpesviridae. Percentage (a) and total number (b) of predicted associations across mammalian orders. Orange bars represent the medians of expected percentages across ensemble members, with uncertainty bars indicating the 5th to 95th percentile range. Blue bars indicate the observed percentages.



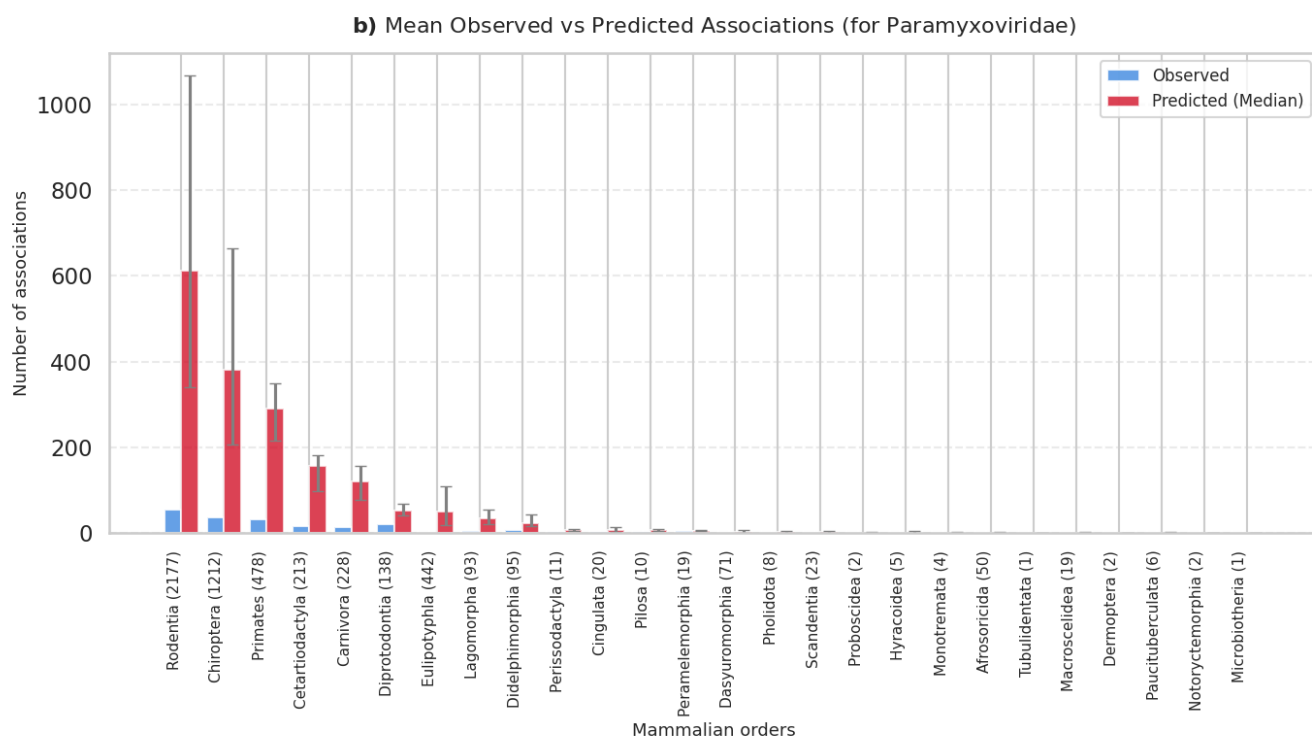
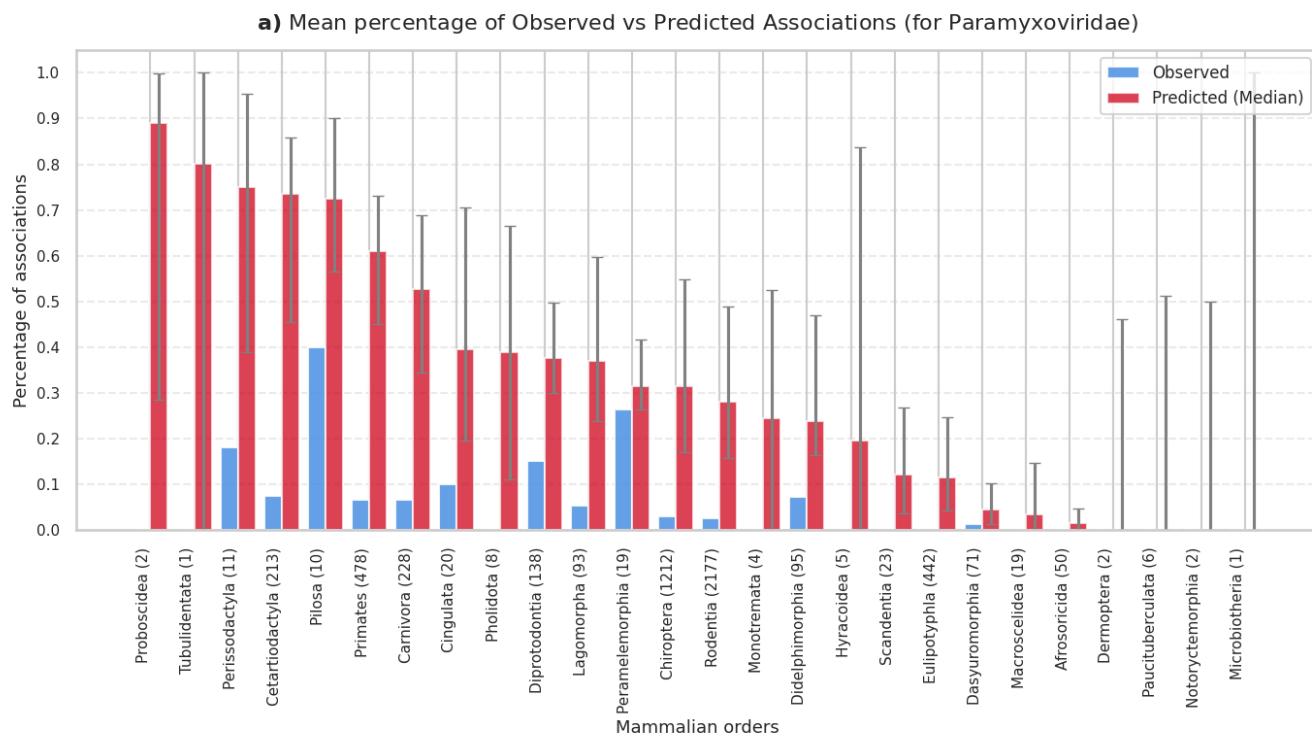
Supplementary Figure SR43 | Predicted associations by mammalian order for Nairoviridae. Percentage (a) and total number (b) of predicted associations across mammalian orders. Orange bars represent the medians of expected percentages across ensemble members, with uncertainty bars indicating the 5th to 95th percentile range. Blue bars indicate the observed percentages.



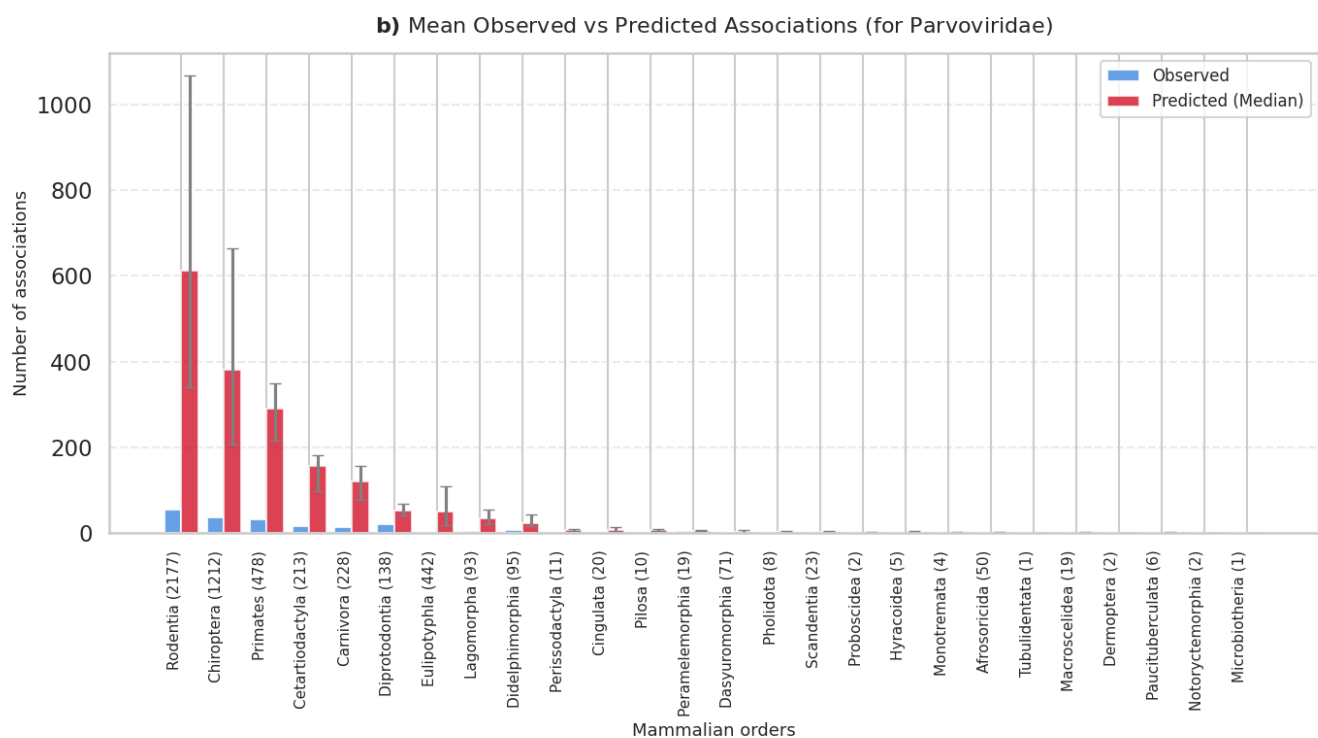
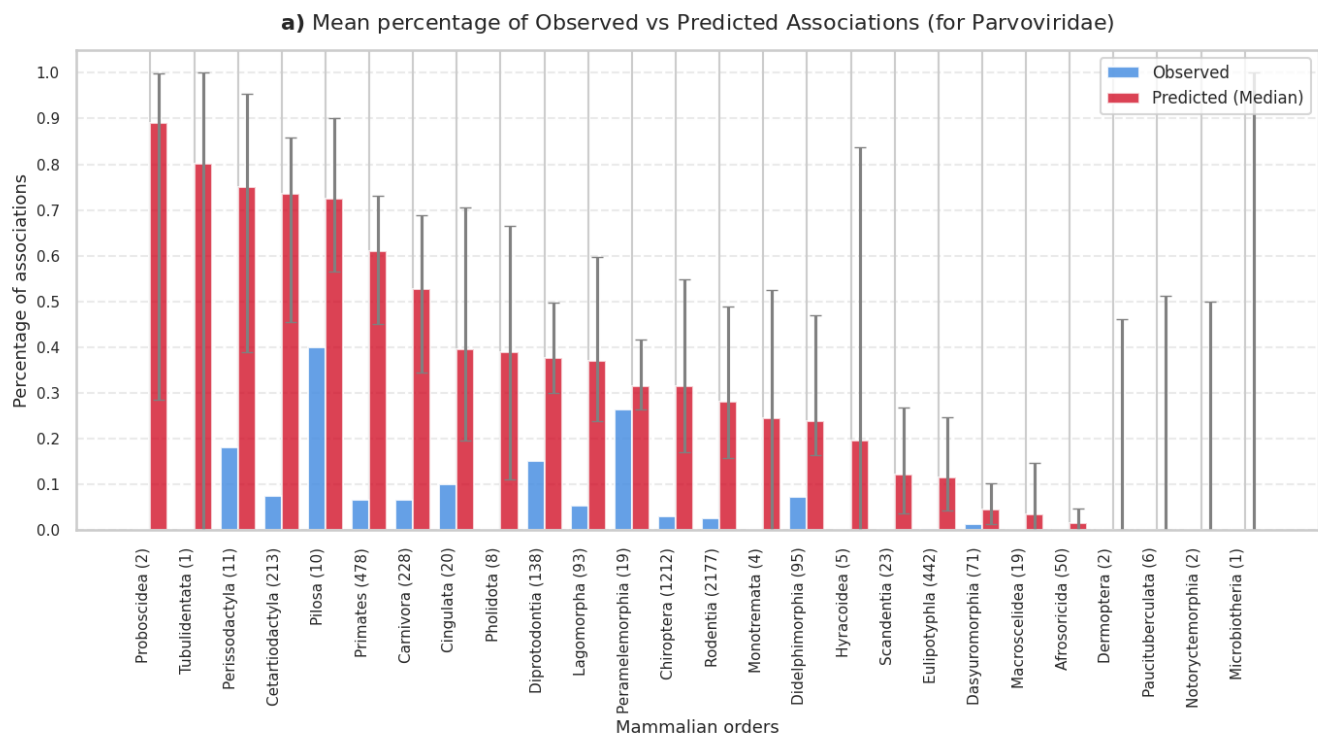
Supplementary Figure SR44 | Predicted associations by mammalian order for Orthomyxoviridae. Percentage (a) and total number (b) of predicted associations across mammalian orders. Orange bars represent the medians of expected percentages across ensemble members, with uncertainty bars indicating the 5th to 95th percentile range. Blue bars indicate the observed percentages.



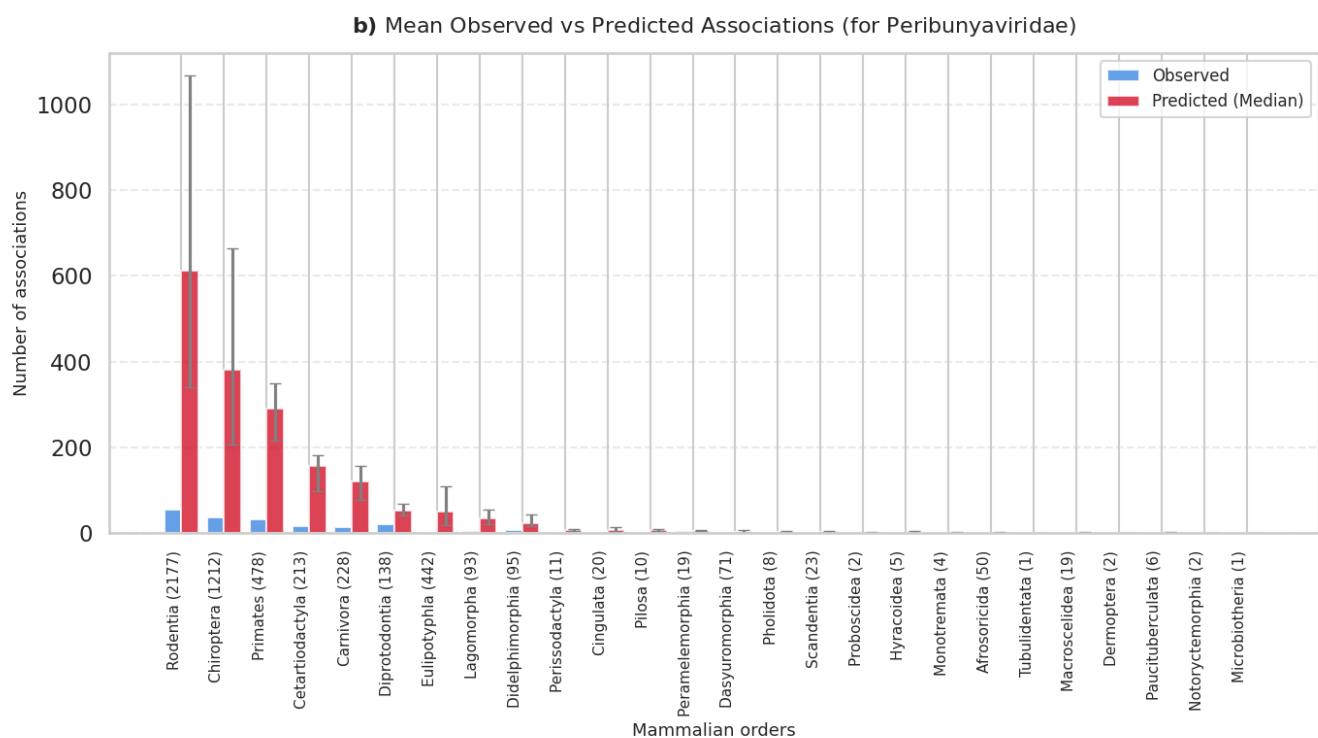
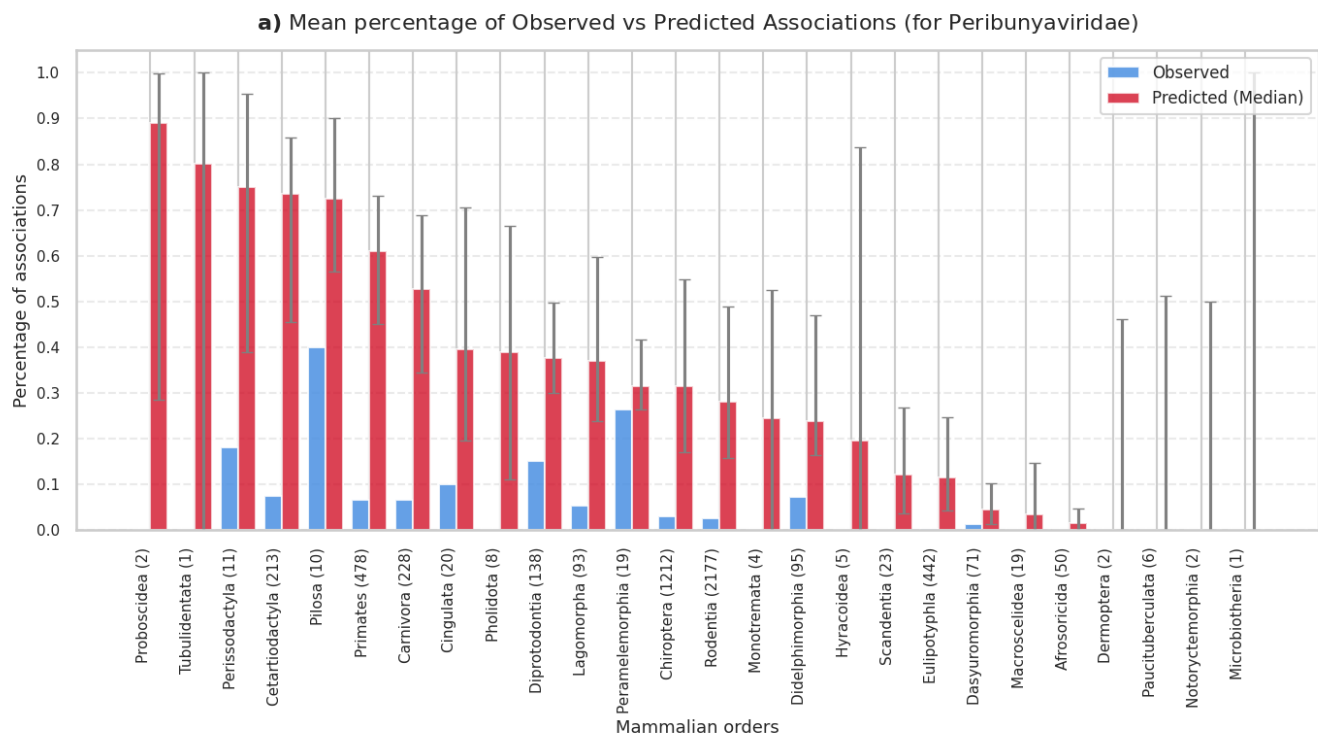
Supplementary Figure SR45 | Predicted associations by mammalian order for Papillomaviridae. Percentage (a) and total number (b) of predicted associations across mammalian orders. Orange bars represent the medians of expected percentages across ensemble members, with uncertainty bars indicating the 5th to 95th percentile range. Blue bars indicate the observed percentages.



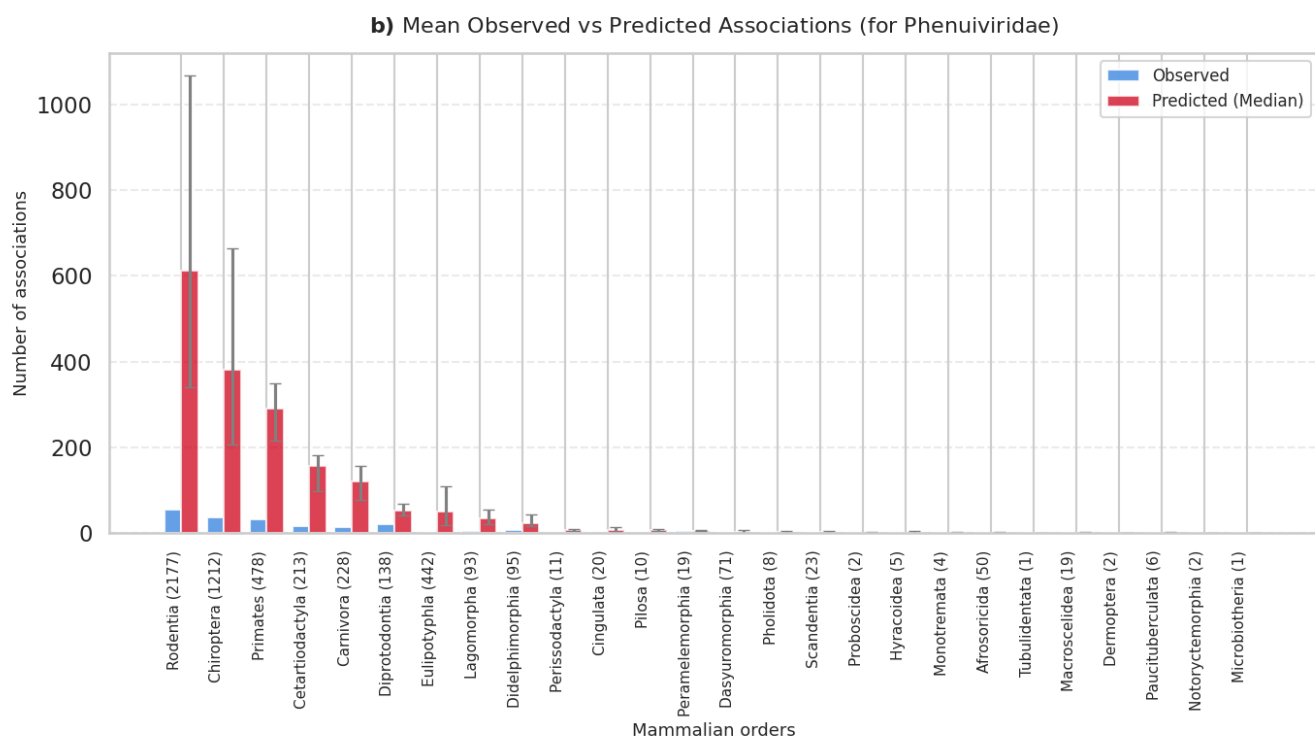
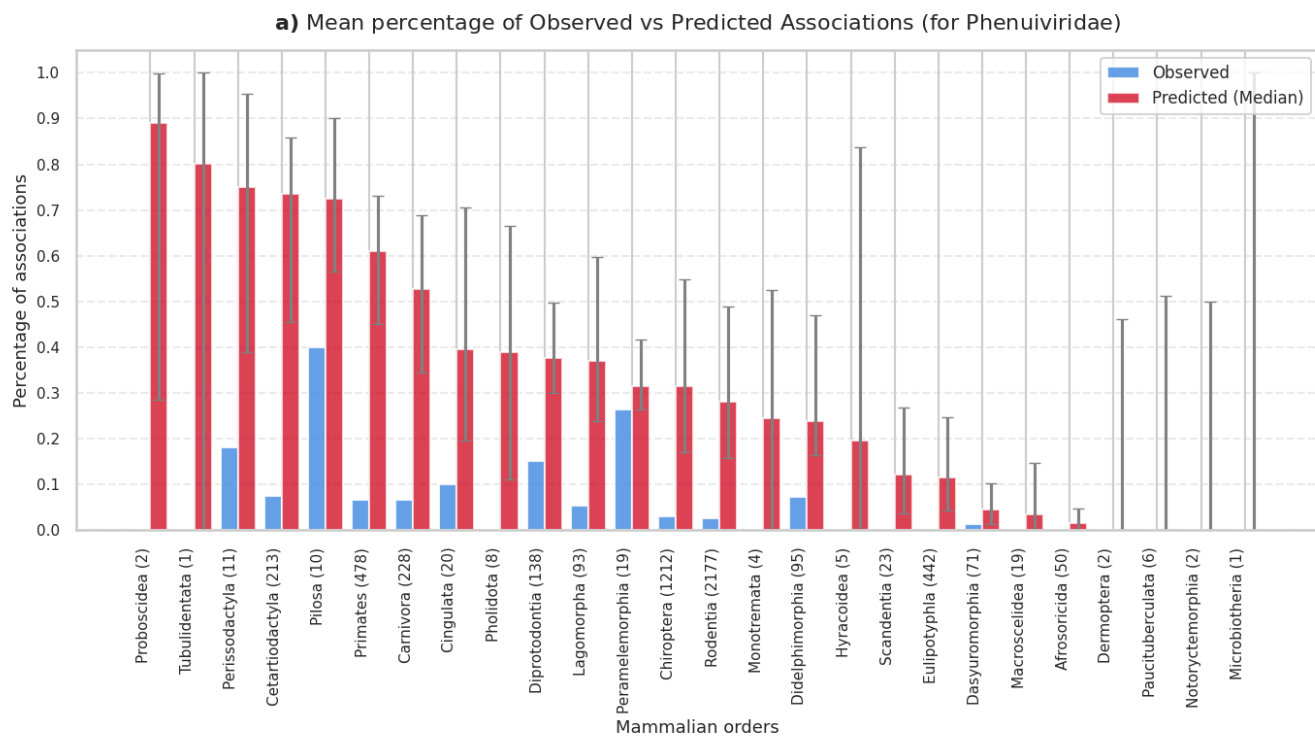
Supplementary Figure SR46 | Predicted associations by mammalian order for Paramyxoviridae. Percentage (a) and total number (b) of predicted associations across mammalian orders. Orange bars represent the medians of expected percentages across ensemble members, with uncertainty bars indicating the 5th to 95th percentile range. Blue bars indicate the observed percentages.



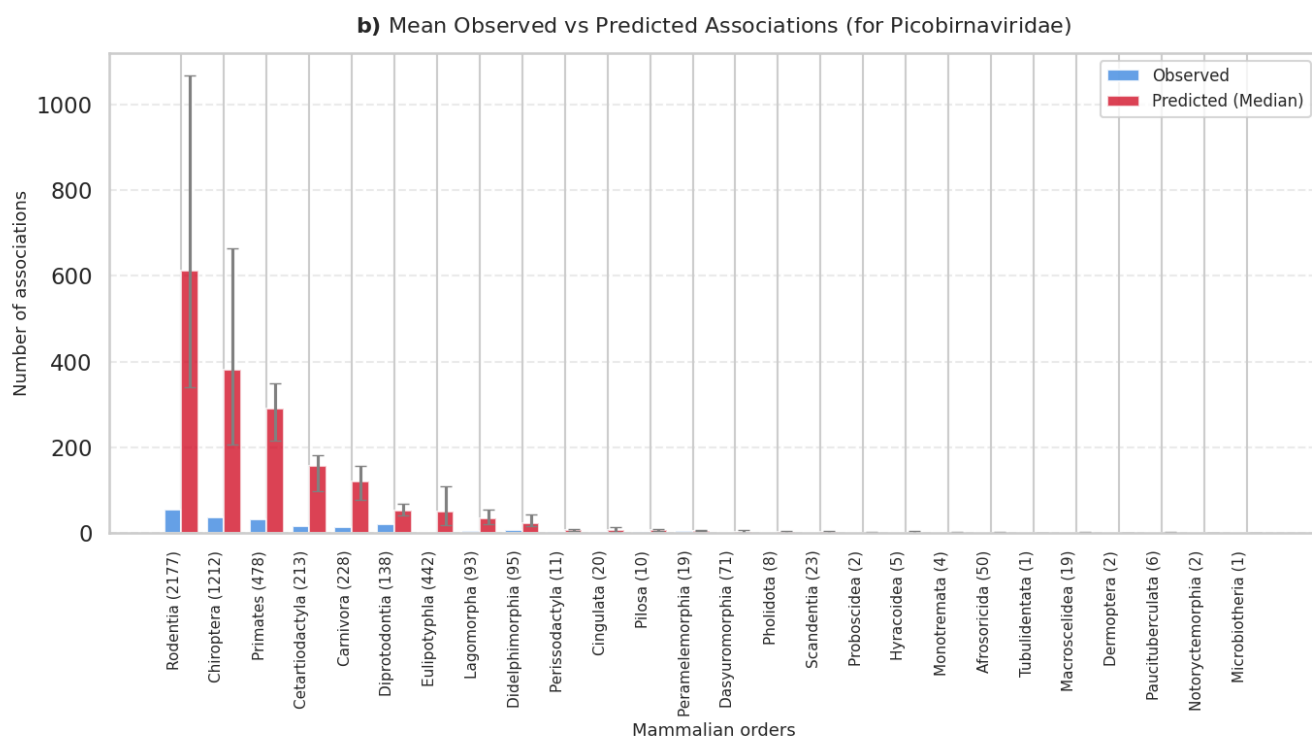
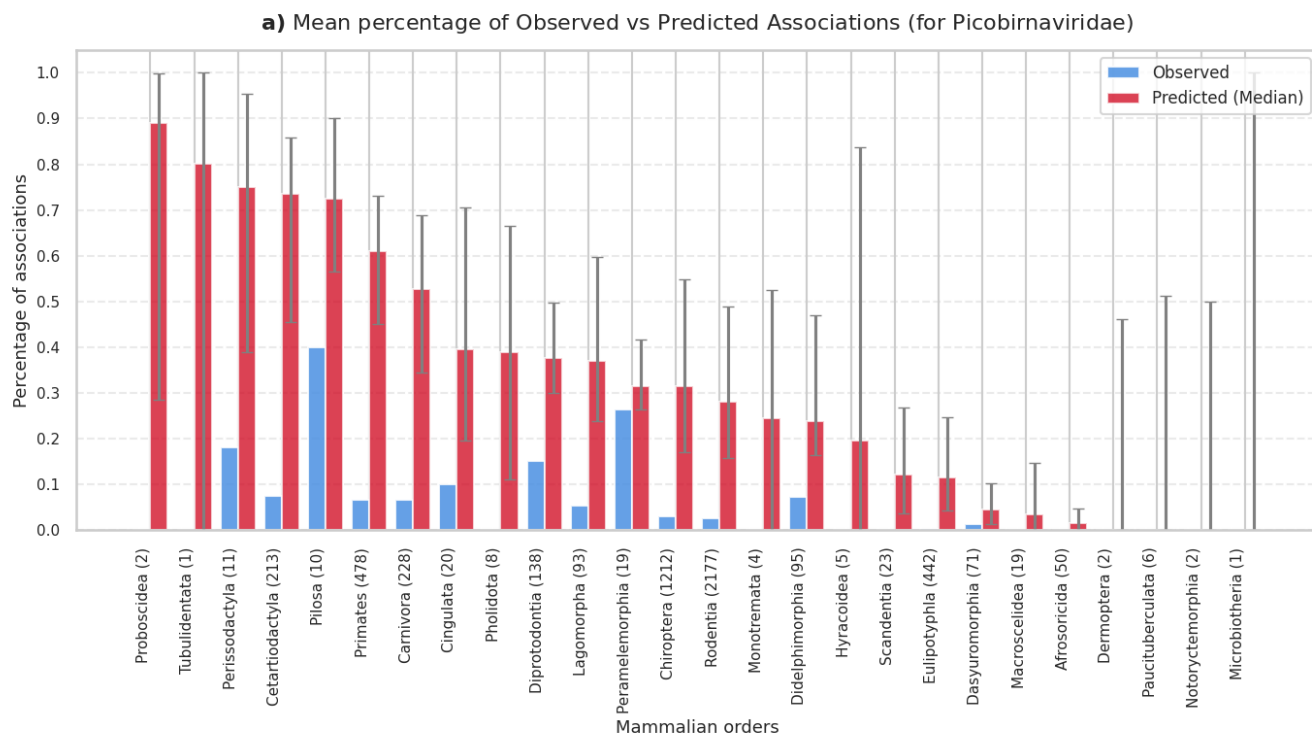
Supplementary Figure SR47 | Predicted associations by mammalian order for Parvoviridae. Percentage (a) and total number (b) of predicted associations across mammalian orders. Orange bars represent the medians of expected percentages across ensemble members, with uncertainty bars indicating the 5th to 95th percentile range. Blue bars indicate the observed percentages.



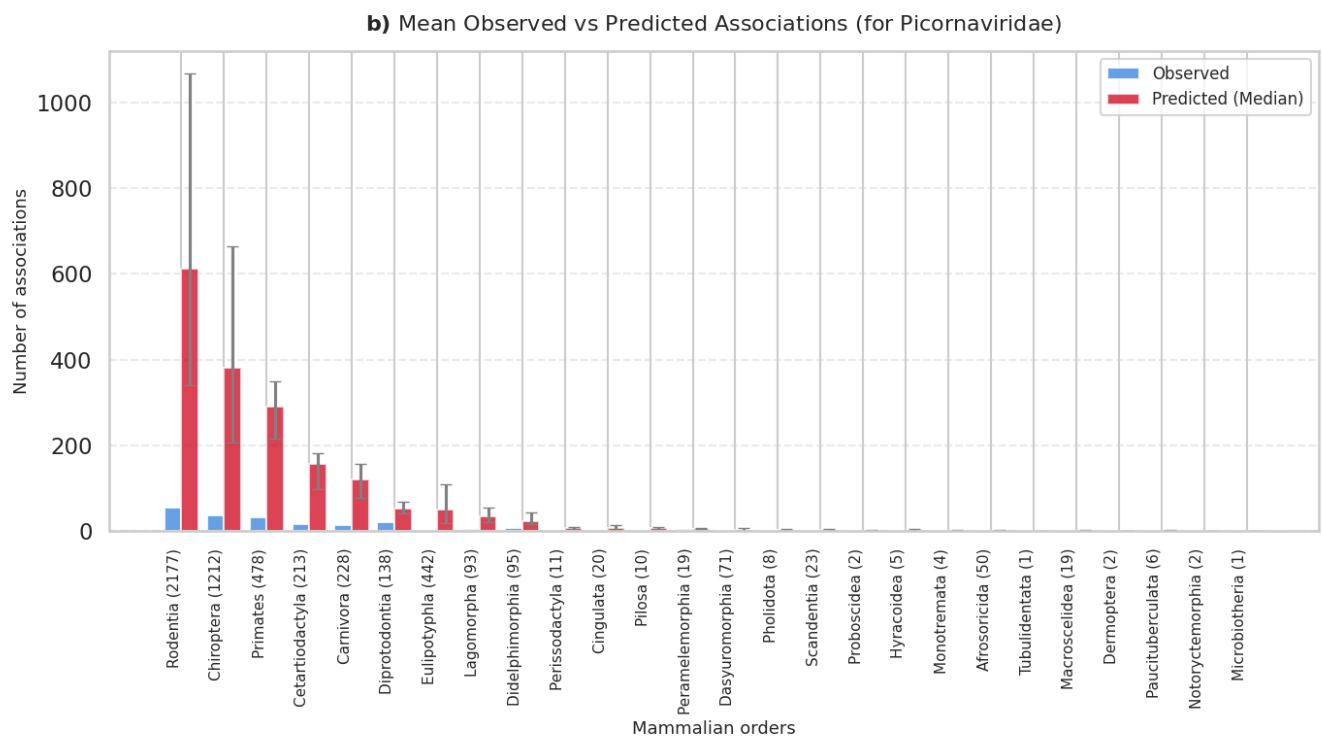
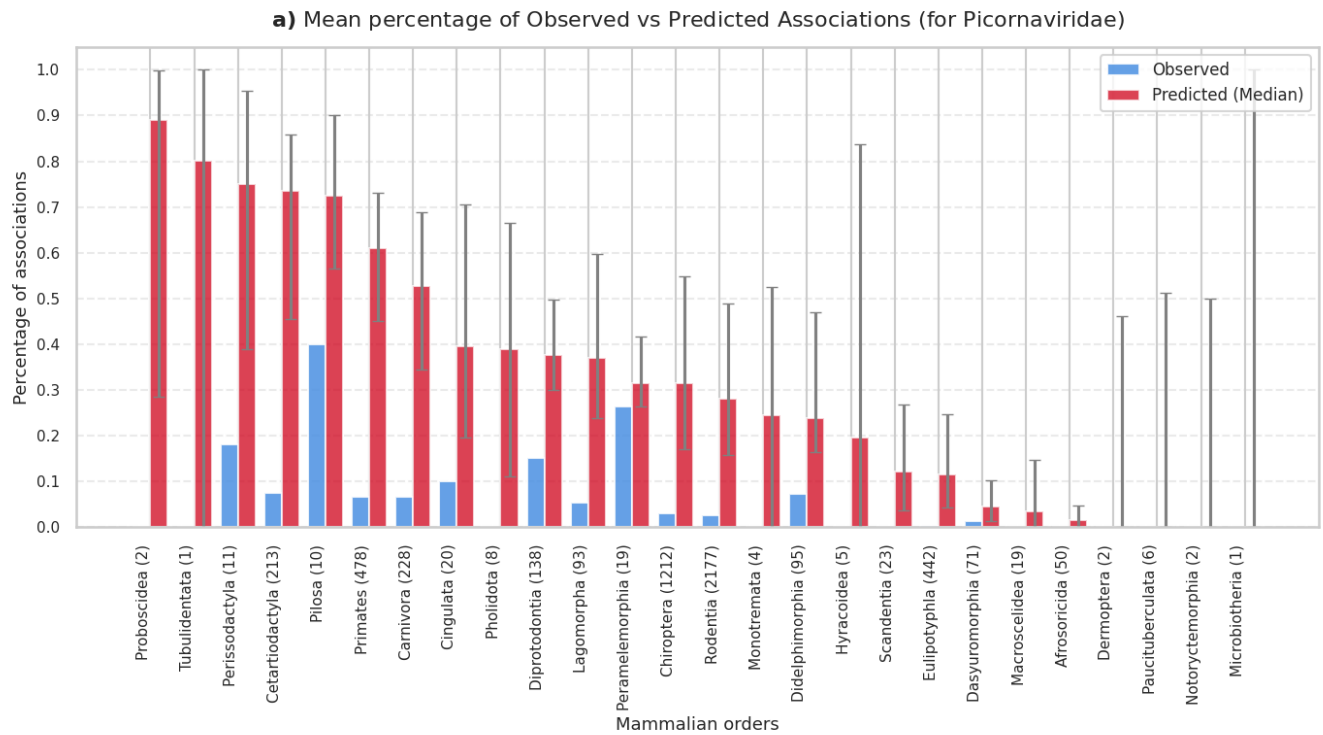
Supplementary Figure SR48 | Predicted associations by mammalian order for Peribunyaviridae. Percentage (a) and total number (b) of predicted associations across mammalian orders. Orange bars represent the medians of expected percentages across ensemble members, with uncertainty bars indicating the 5th to 95th percentile range. Blue bars indicate the observed percentages.



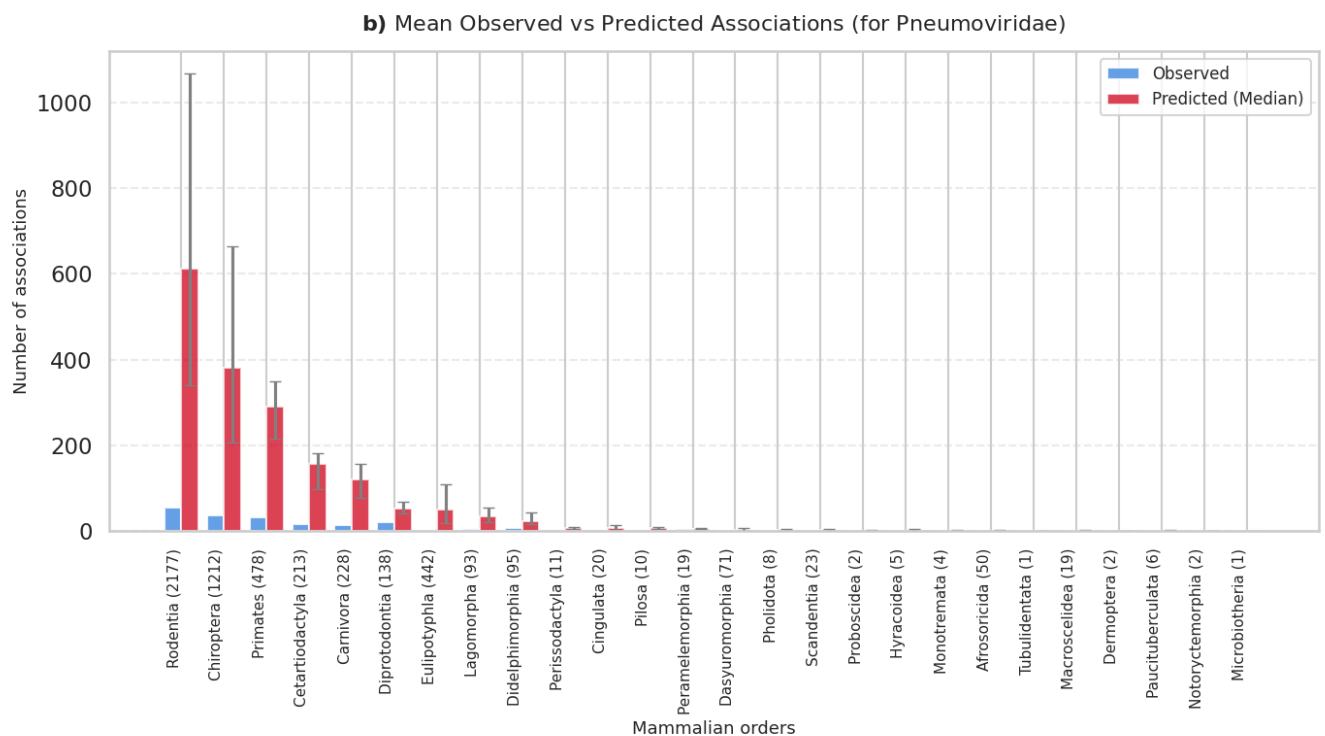
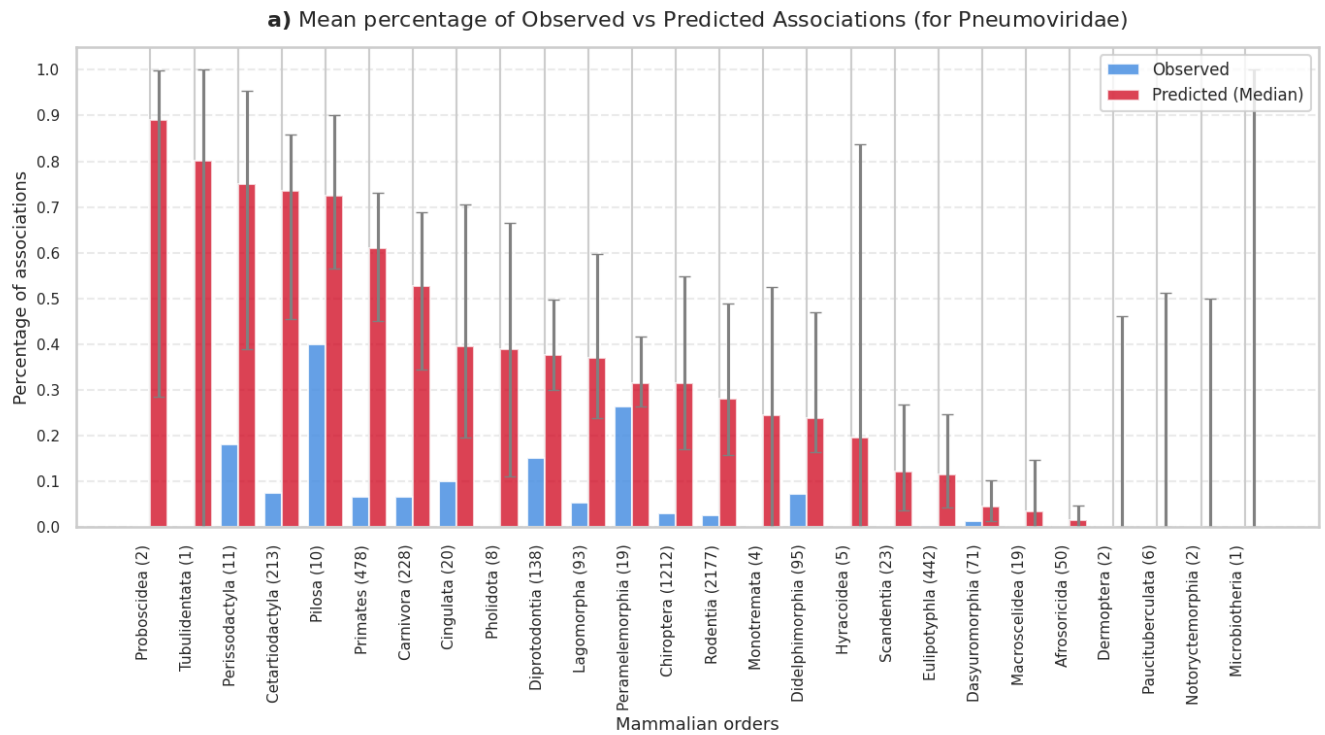
Supplementary Figure SR49 | Predicted associations by mammalian order for Phenuiviridae. Percentage (a) and total number (b) of predicted associations across mammalian orders. Orange bars represent the medians of expected percentages across ensemble members, with uncertainty bars indicating the 5th to 95th percentile range. Blue bars indicate the observed percentages.



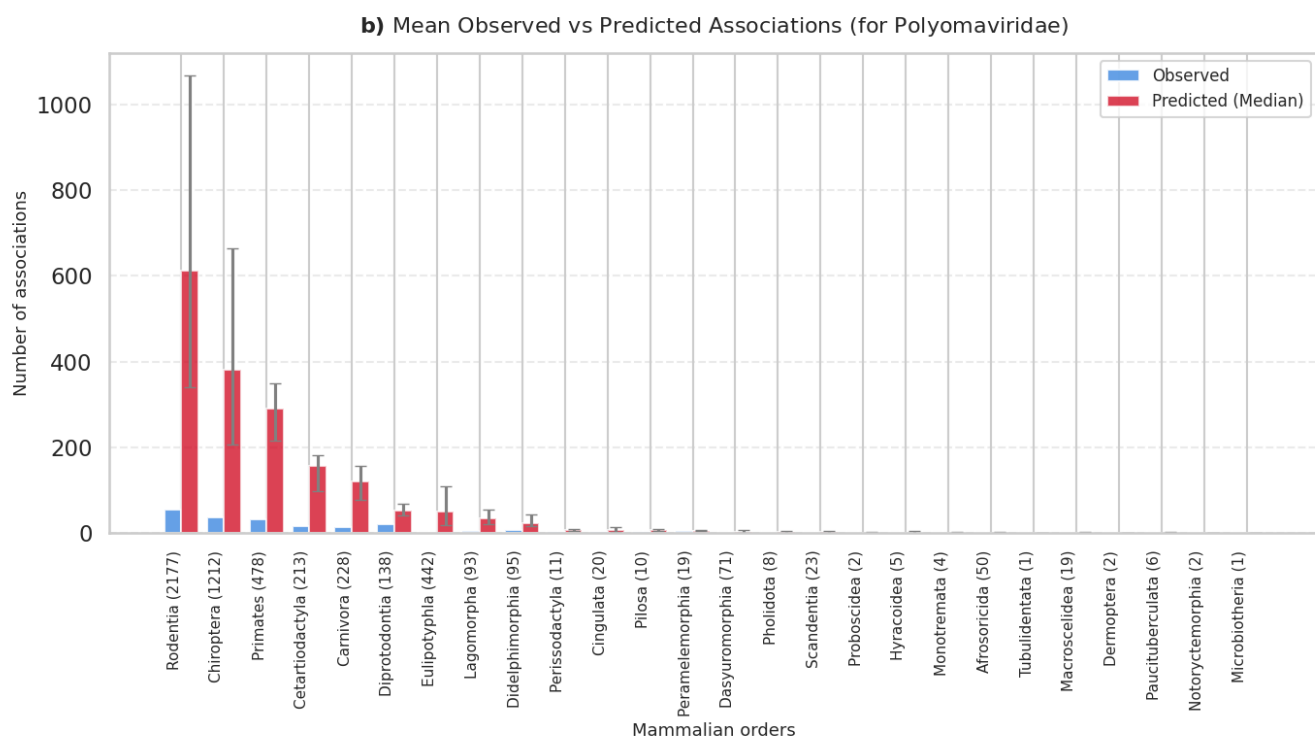
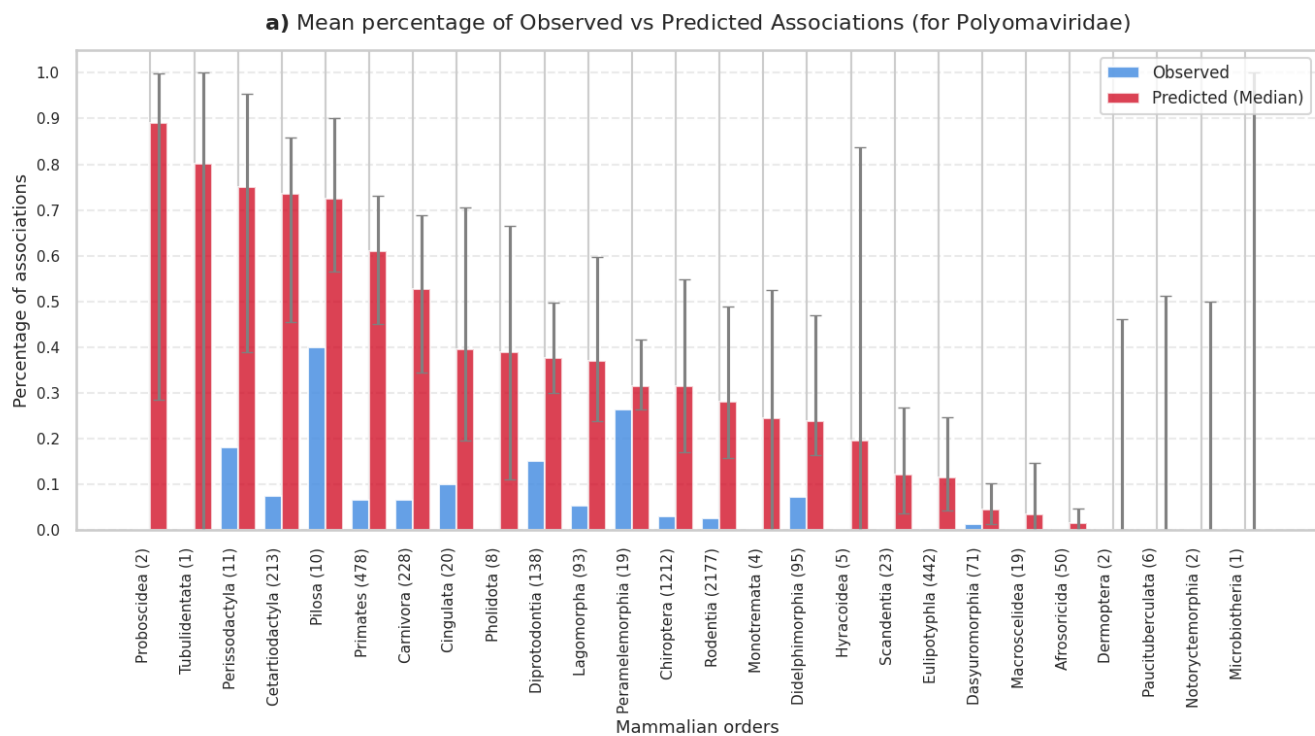
Supplementary Figure SR50 | Predicted associations by mammalian order for Picobirnaviridae. Percentage (a) and total number (b) of predicted associations across mammalian orders. Orange bars represent the medians of expected percentages across ensemble members, with uncertainty bars indicating the 5th to 95th percentile range. Blue bars indicate the observed percentages.



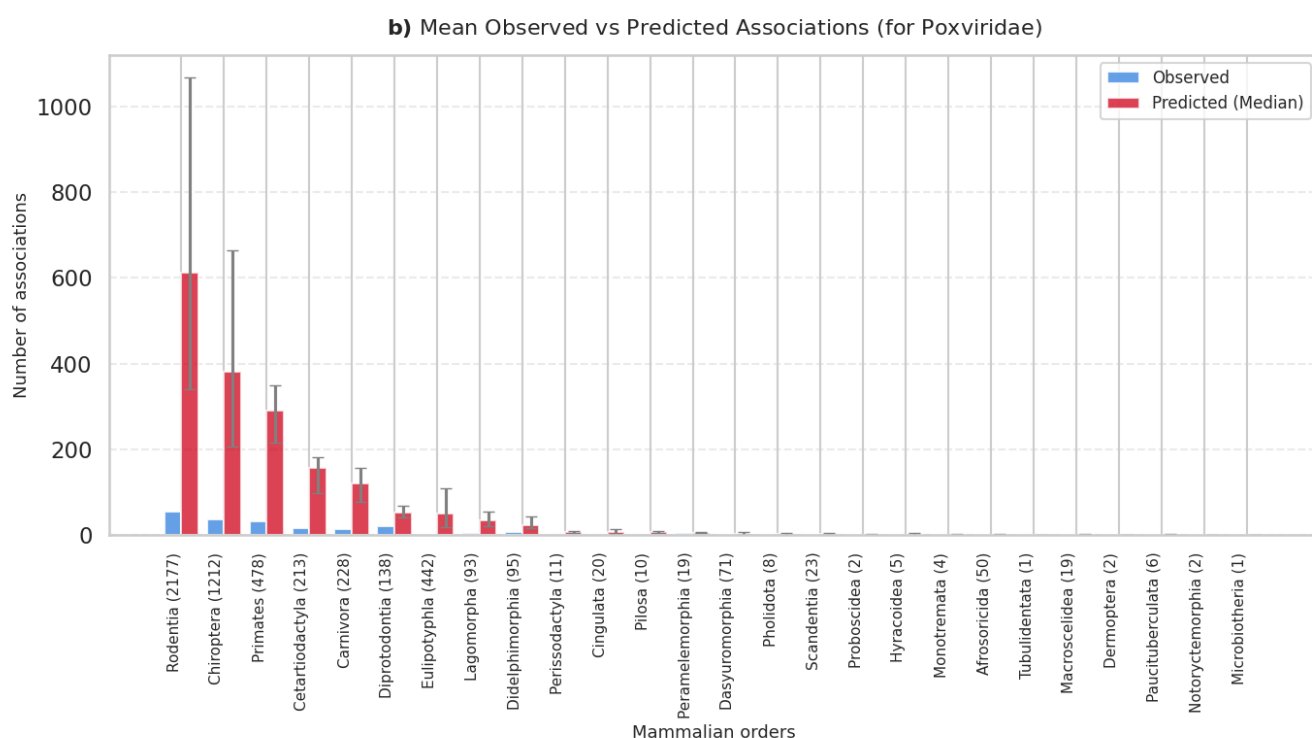
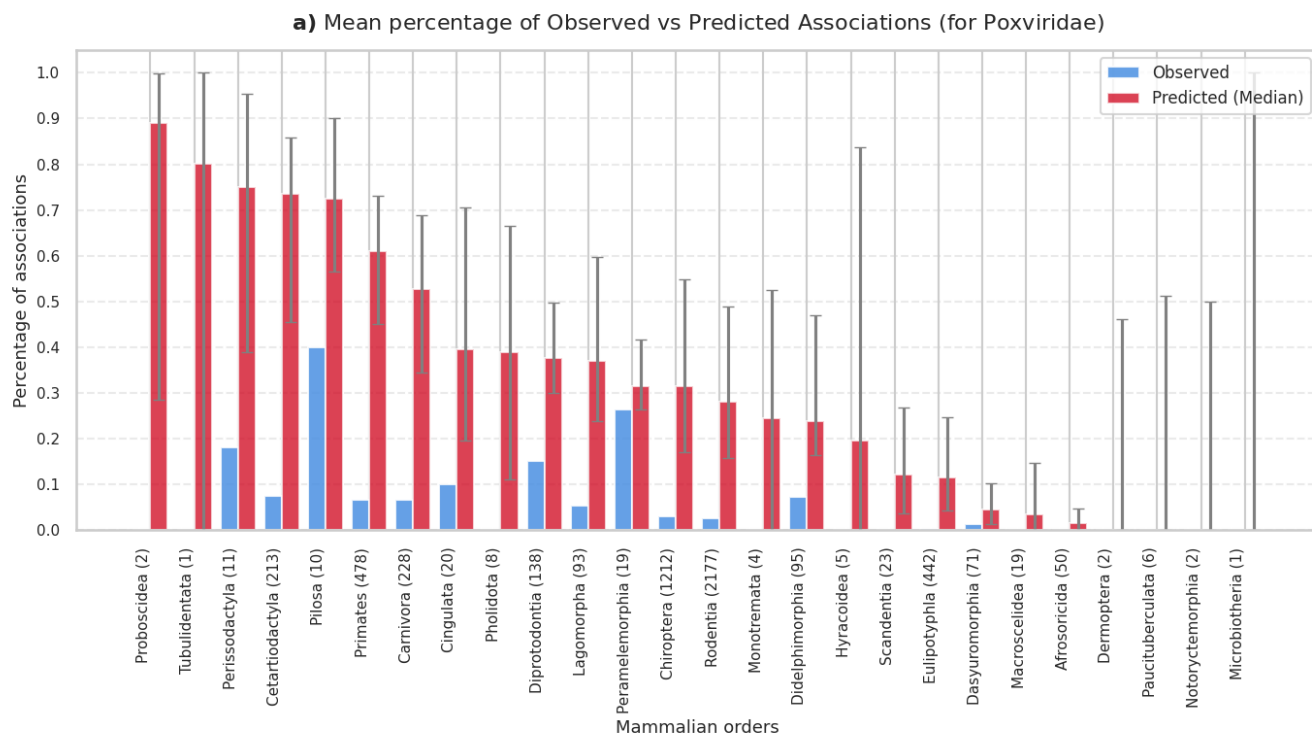
Supplementary Figure SR51 | Predicted associations by mammalian order for Picornaviridae. Percentage (a) and total number (b) of predicted associations across mammalian orders. Orange bars represent the medians of expected percentages across ensemble members, with uncertainty bars indicating the 5th to 95th percentile range. Blue bars indicate the observed percentages.



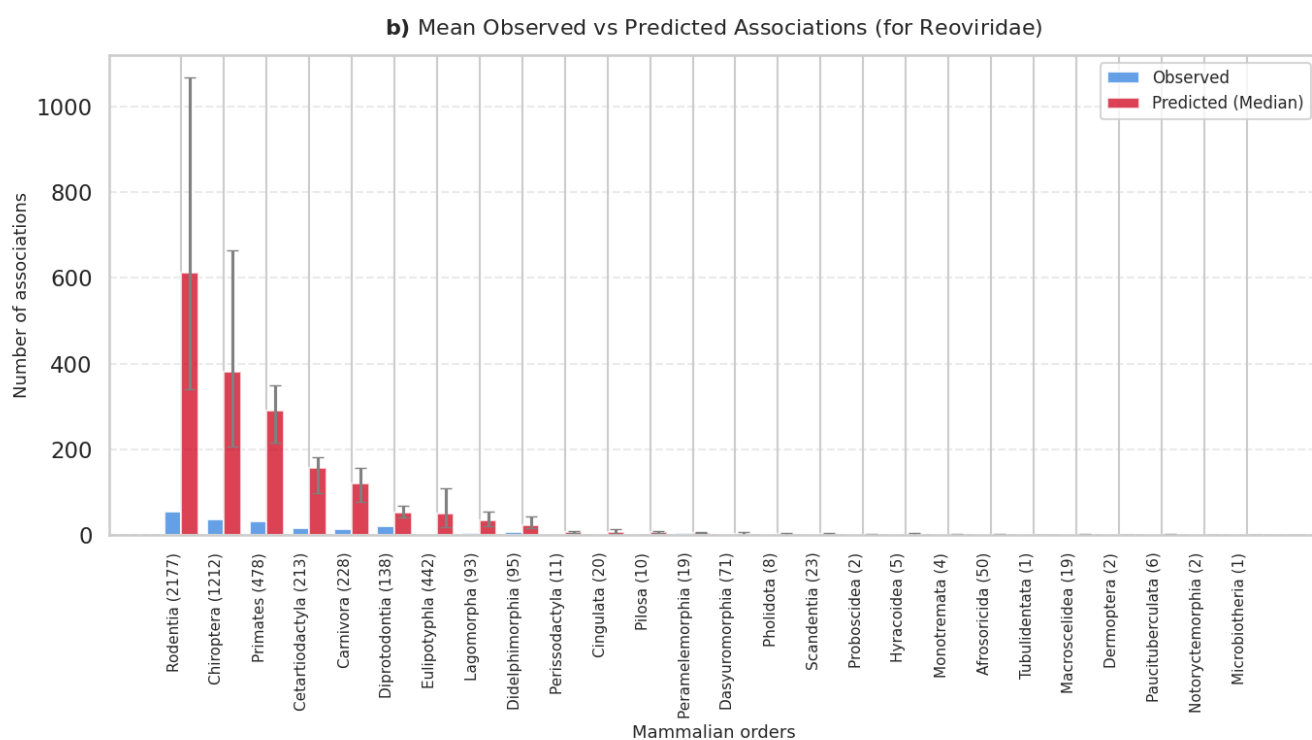
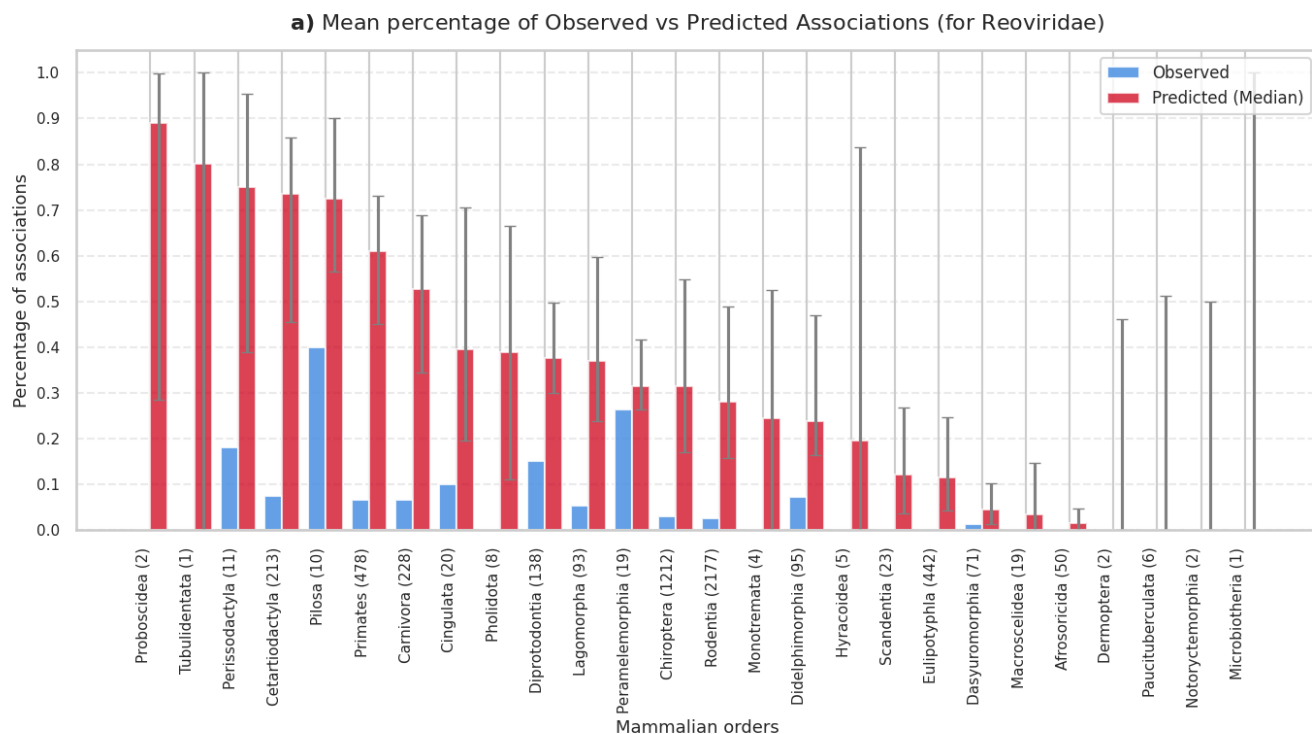
Supplementary Figure SR52 | Predicted associations by mammalian order for Pneumoviridae. Percentage (a) and total number (b) of predicted associations across mammalian orders. Orange bars represent the medians of expected percentages across ensemble members, with uncertainty bars indicating the 5th to 95th percentile range. Blue bars indicate the observed percentages.



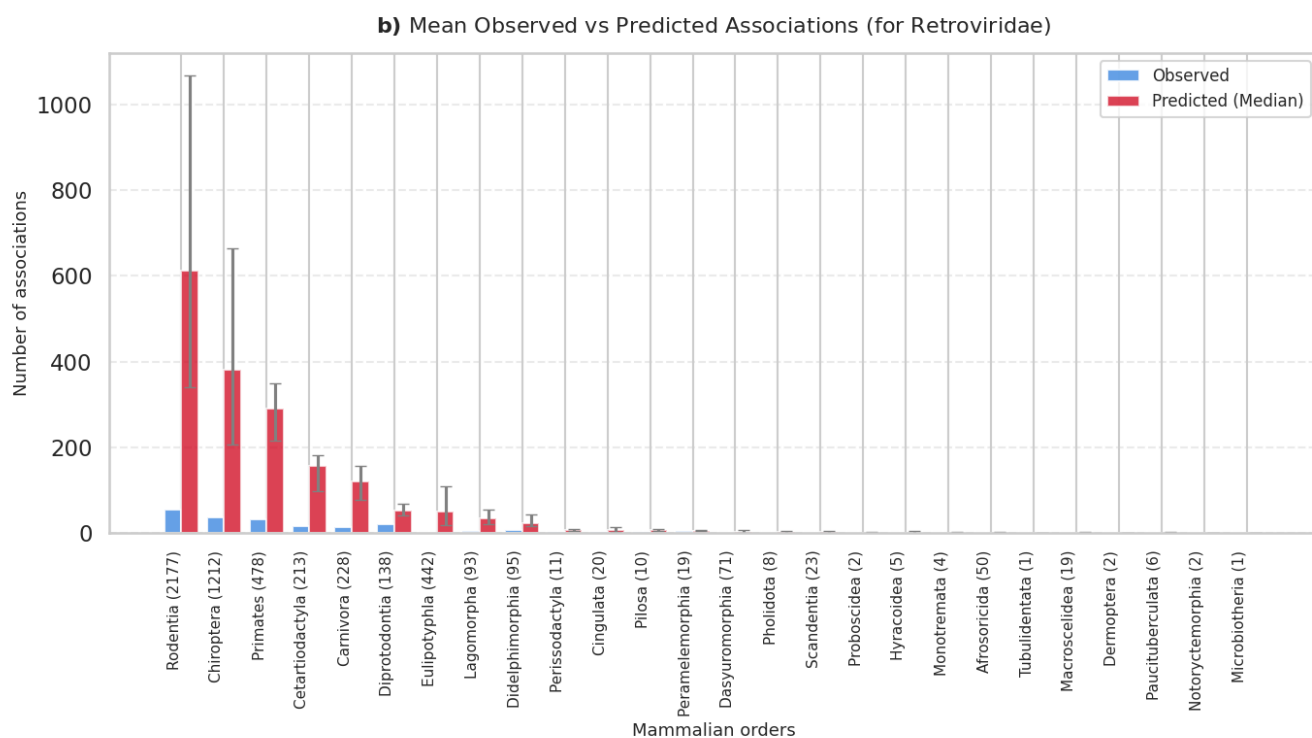
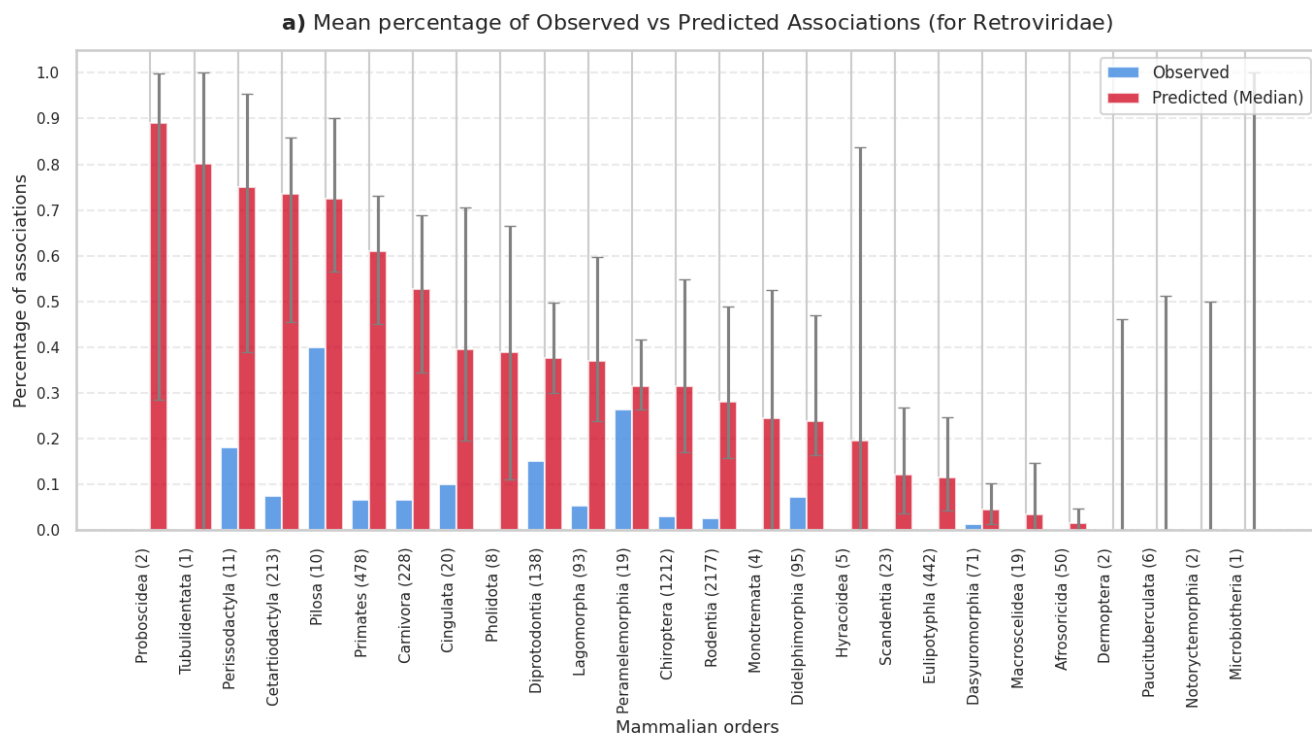
Supplementary Figure SR53 | Predicted associations by mammalian order for Polyomaviridae. Percentage (a) and total number (b) of predicted associations across mammalian orders. Orange bars represent the medians of expected percentages across ensemble members, with uncertainty bars indicating the 5th to 95th percentile range. Blue bars indicate the observed percentages.



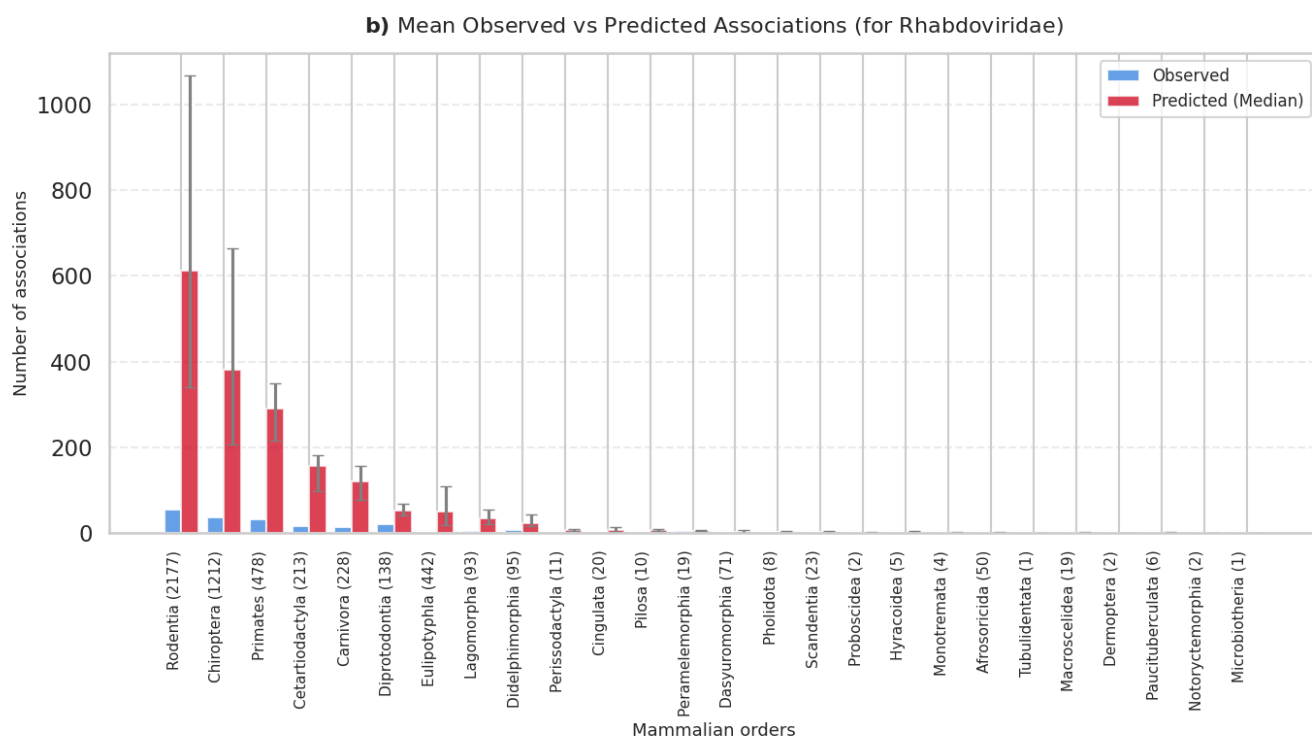
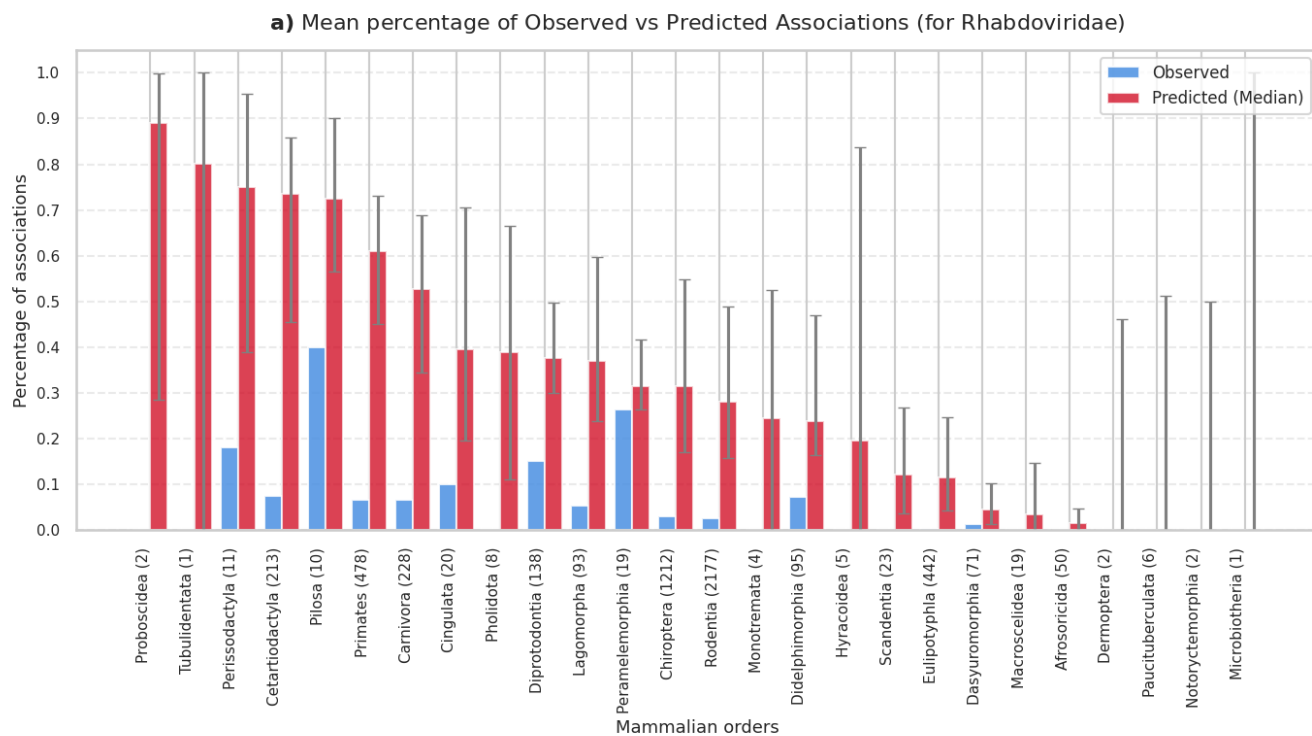
Supplementary Figure SR54 | Predicted associations by mammalian order for Poxviridae. Percentage (a) and total number (b) of predicted associations across mammalian orders. Orange bars represent the medians of expected percentages across ensemble members, with uncertainty bars indicating the 5th to 95th percentile range. Blue bars indicate the observed percentages.



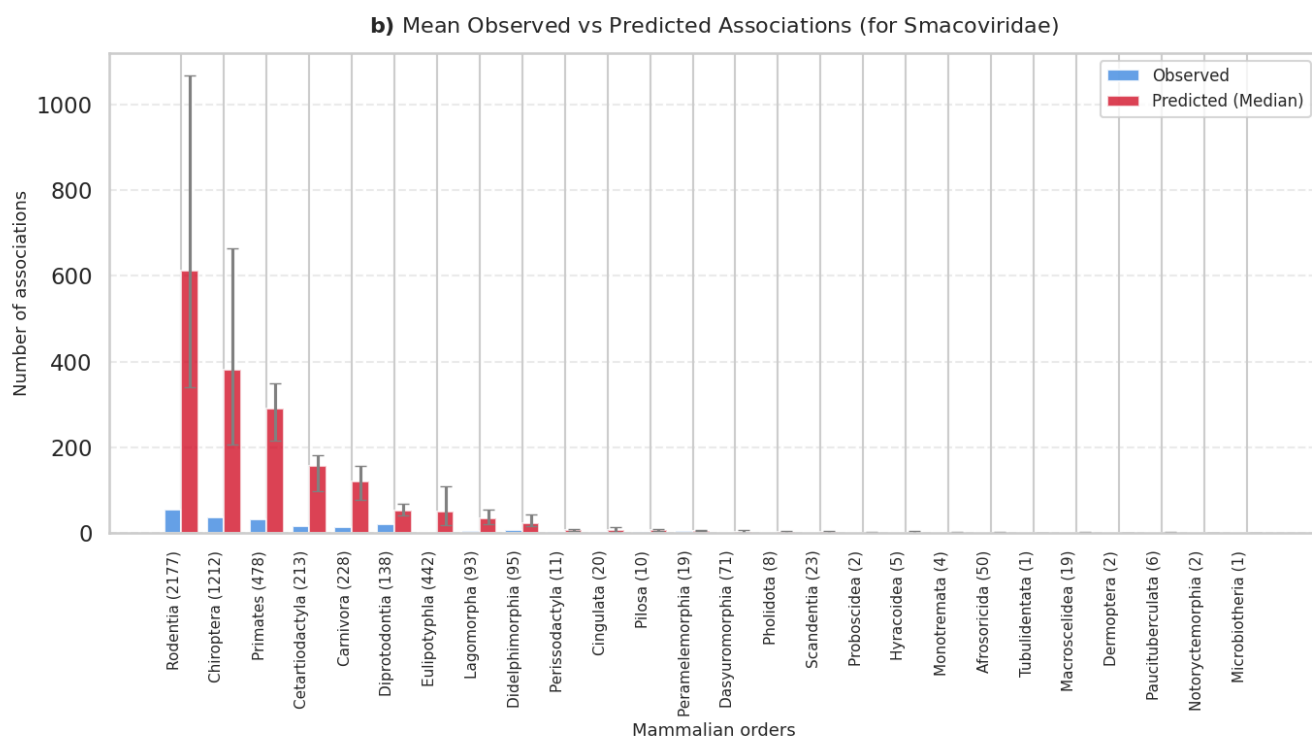
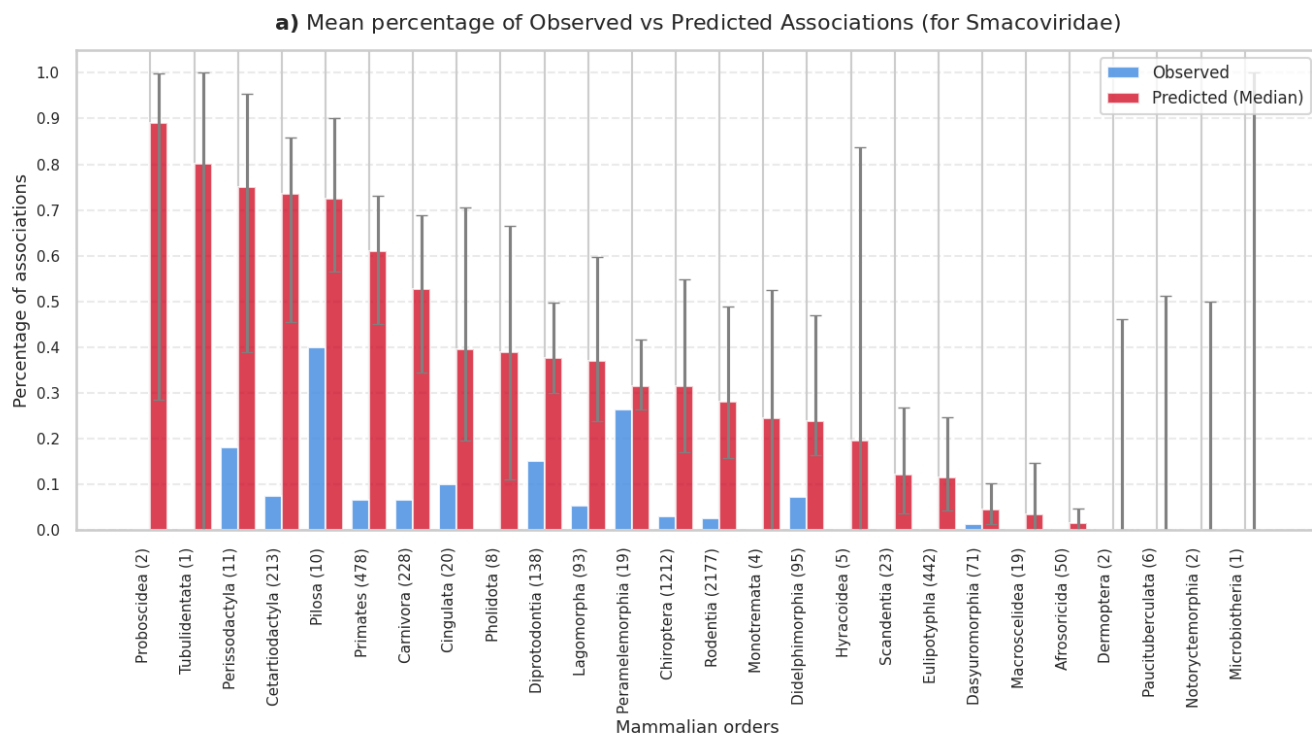
Supplementary Figure SR55 | Predicted associations by mammalian order for Reoviridae. Percentage (a) and total number (b) of predicted associations across mammalian orders. Orange bars represent the medians of expected percentages across ensemble members, with uncertainty bars indicating the 5th to 95th percentile range. Blue bars indicate the observed percentages.



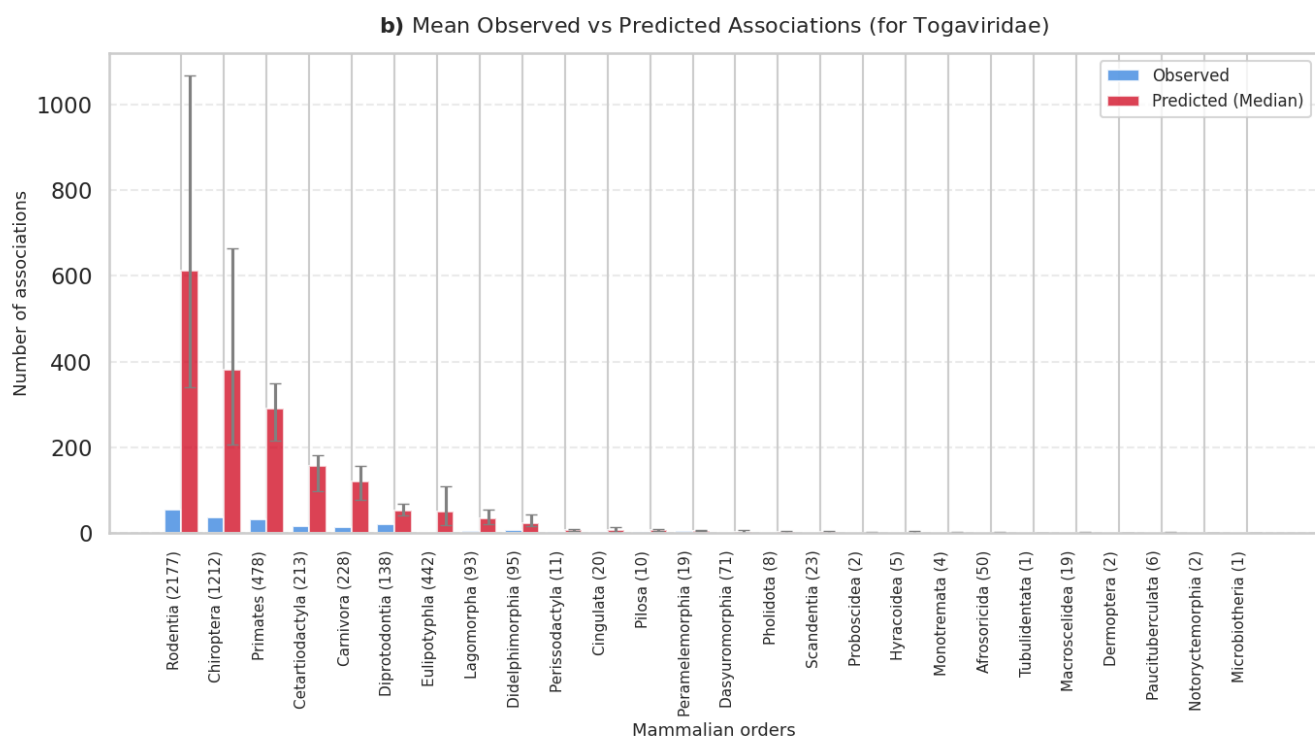
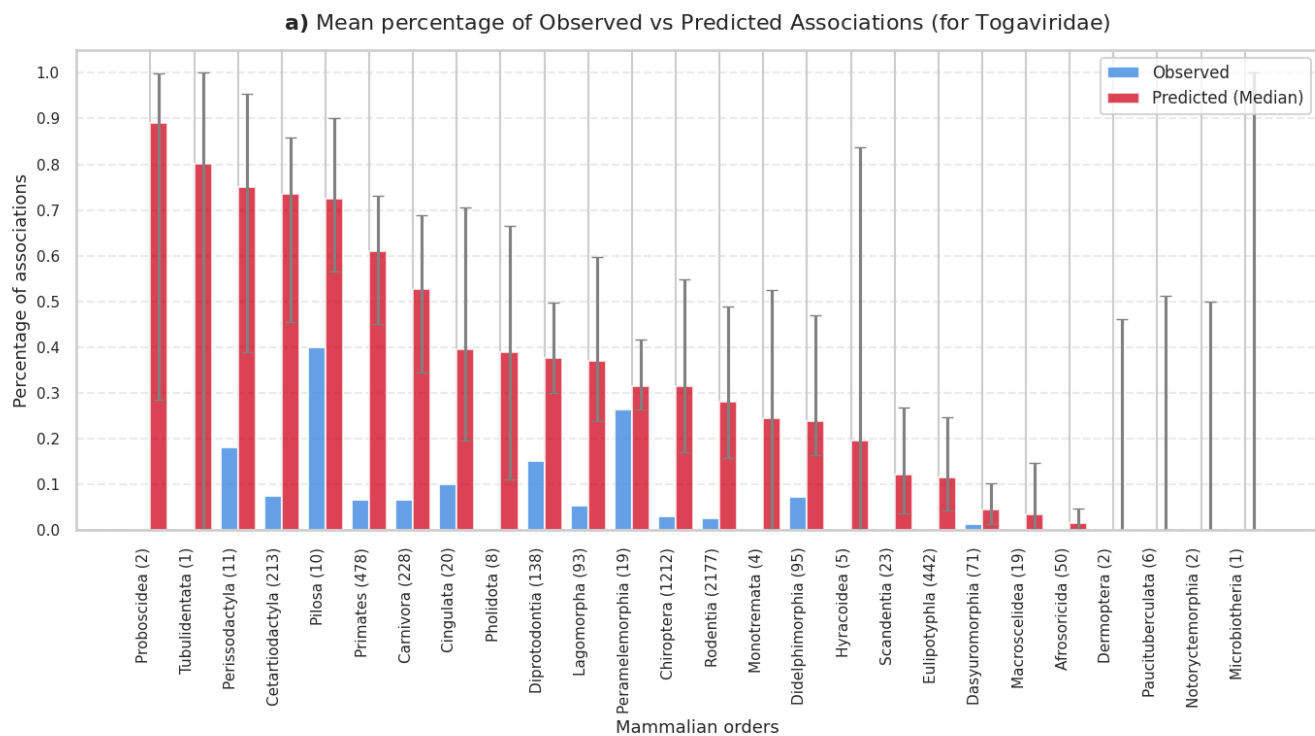
Supplementary Figure SR56 | Predicted associations by mammalian order for Retroviridae. Percentage (a) and total number (b) of predicted associations across mammalian orders. Orange bars represent the medians of expected percentages across ensemble members, with uncertainty bars indicating the 5th to 95th percentile range. Blue bars indicate the observed percentages.



Supplementary Figure SR57 | Predicted associations by mammalian order for Rhabdoviridae. Percentage (a) and total number (b) of predicted associations across mammalian orders. Orange bars represent the medians of expected percentages across ensemble members, with uncertainty bars indicating the 5th to 95th percentile range. Blue bars indicate the observed percentages.



Supplementary Figure SR58 | Predicted associations by mammalian order for Smacoviridae. Percentage (a) and total number (b) of predicted associations across mammalian orders. Orange bars represent the medians of expected percentages across ensemble members, with uncertainty bars indicating the 5th to 95th percentile range. Blue bars indicate the observed percentages.

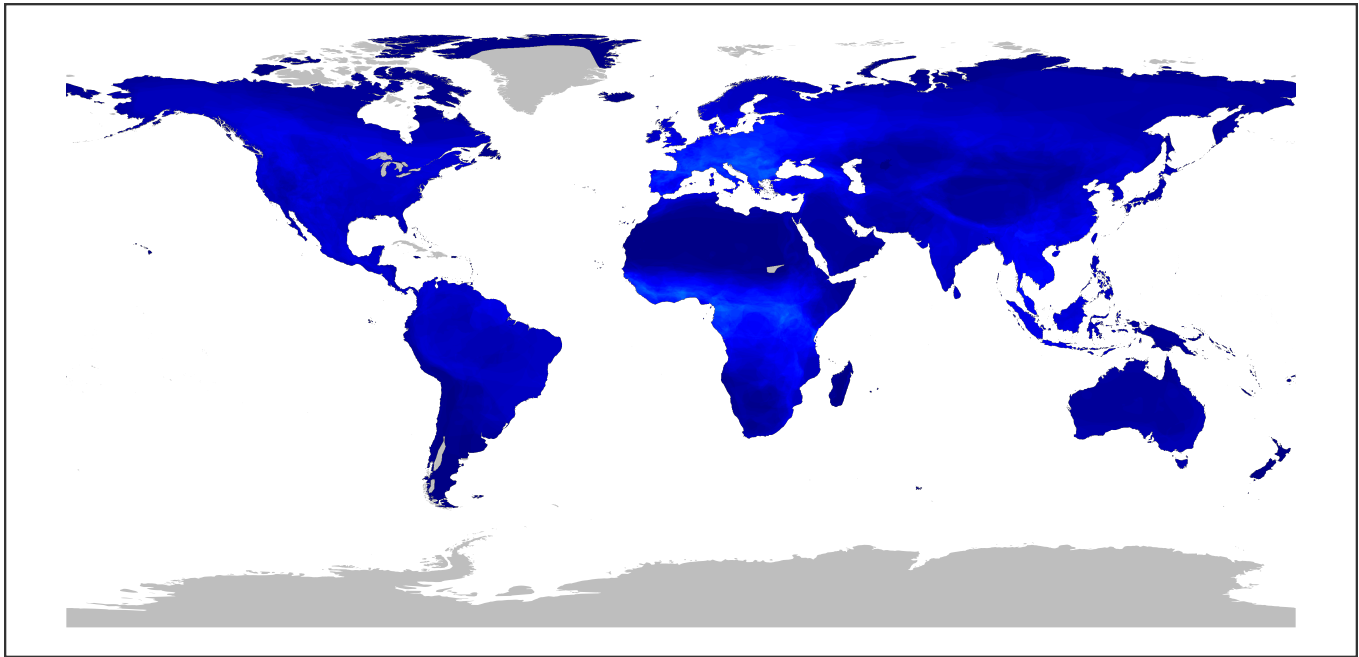


Supplementary Figure SR59 | Predicted associations by mammalian order for Togaviridae. Percentage (a) and total number (b) of predicted associations across mammalian orders. Orange bars represent the medians of expected percentages across ensemble members, with uncertainty bars indicating the 5th to 95th percentile range. Blue bars indicate the observed percentages.


2.2 Predicted geographic distribution of the hosts

Here, for each viral family, we show the observed vs predicted geographic distribution of the hosts (Supplementary Figures SR60–SR92).

a) Observed distribution map for Adenoviridae

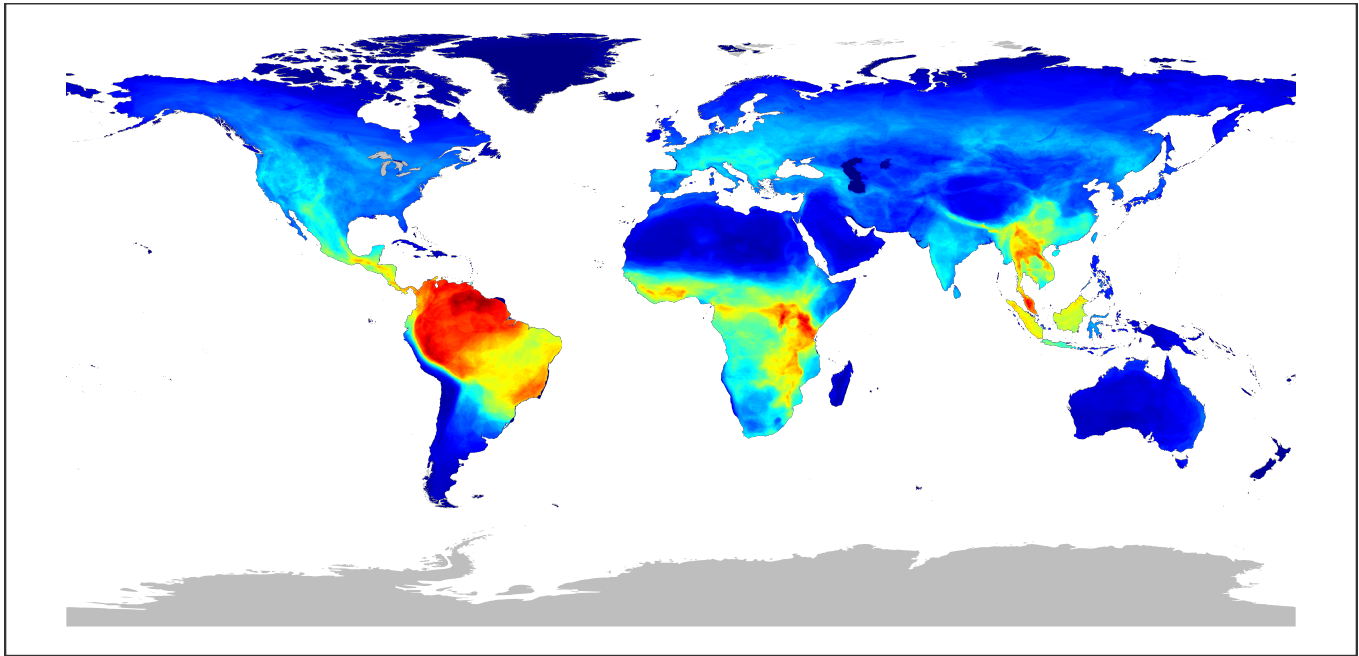


Number of observed susceptible species




0 50 100 150

b) Predicted distribution map for Adenoviridae ($p > 0.5$)



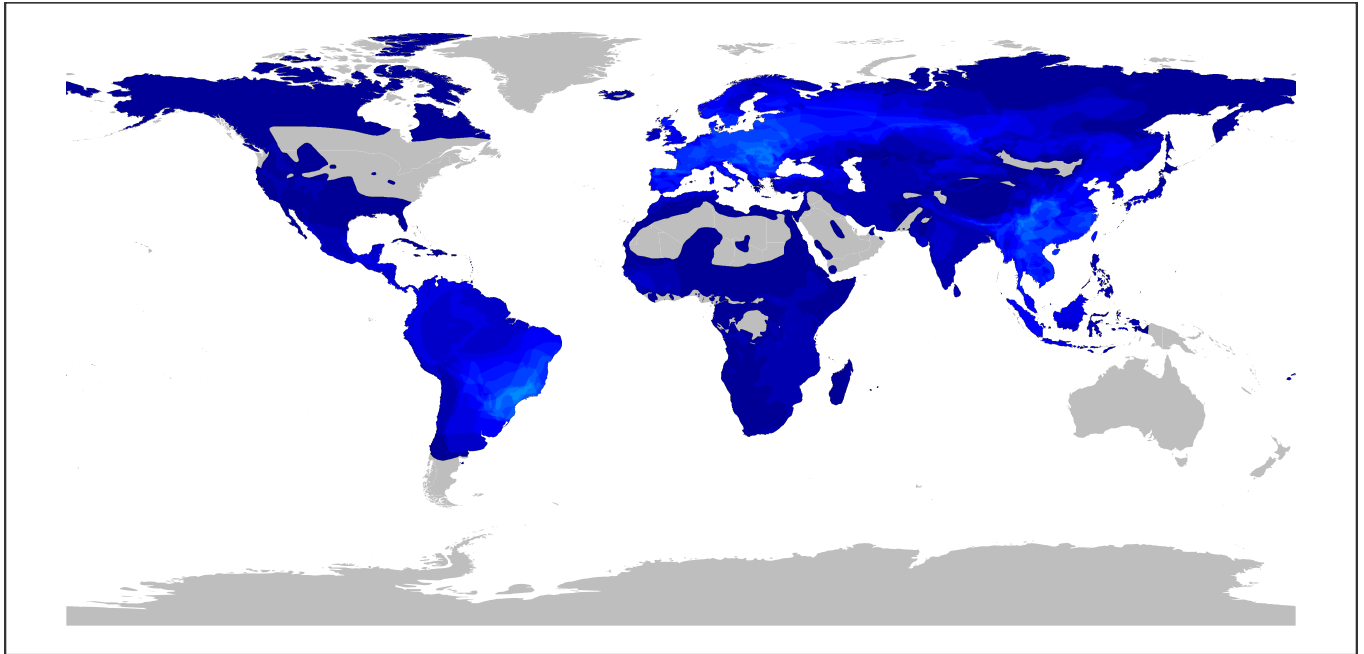
Number of observed and predicted susceptible species




0 50 100 150

Supplementary Figure SR60 | Geographic distribution of associations for Adenoviridae. The panels show the observed (a) and predicted (b) richness of mammal species associated with the viral family. A mammal species was considered predicted to be associated with the viral family if the median predicted probability exceeded 0.5.

a) Observed distribution map for Anelloviridae

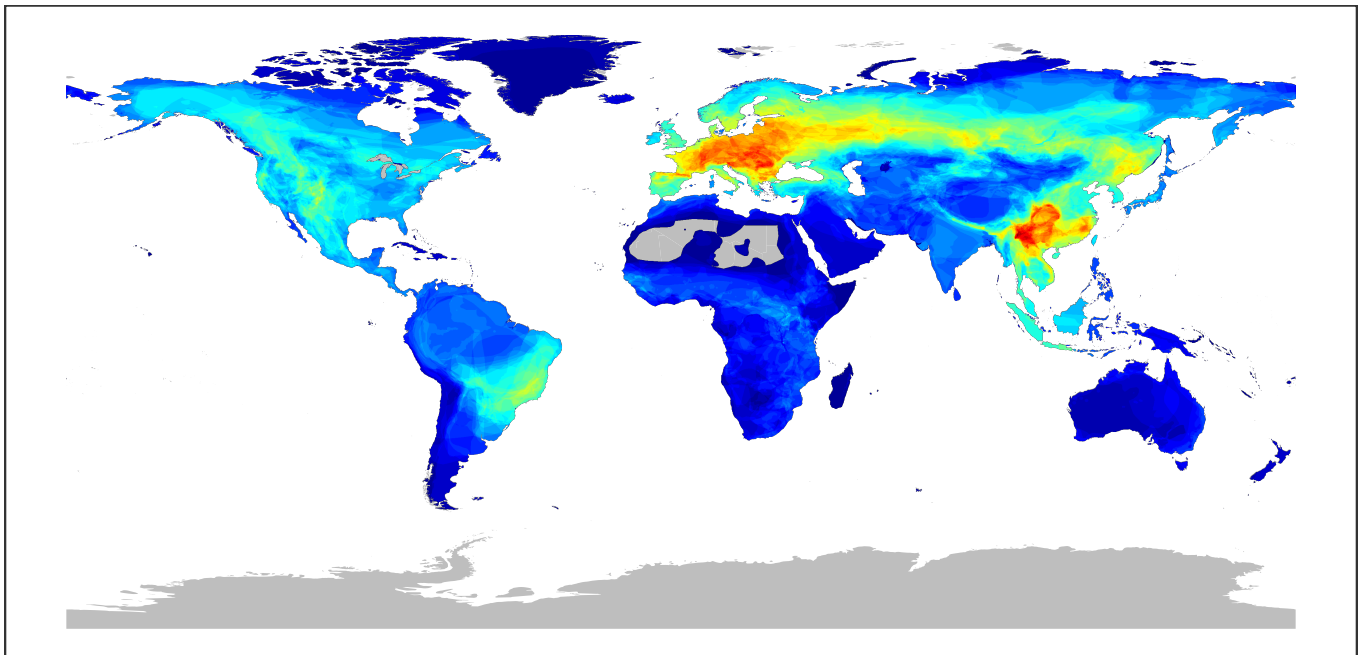


Number of observed susceptible species




0 10 20 30 40

b) Predicted distribution map for Anelloviridae ($p > 0.5$)



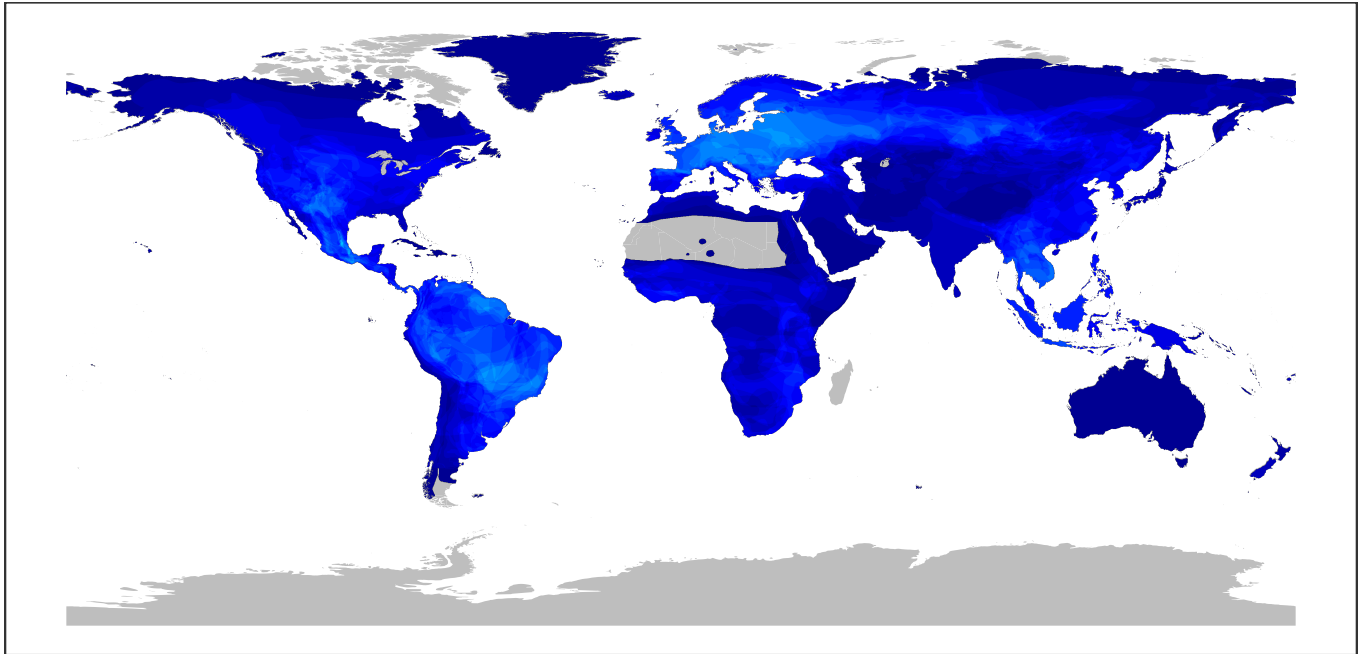
Number of observed and predicted susceptible species




0 10 20 30 40

Supplementary Figure SR61 | Geographic distribution of associations for Anelloviridae. The panels show the observed (a) and predicted (b) richness of mammal species associated with the viral family. A mammal species was considered predicted to be associated with the viral family if the median predicted probability exceeded 0.5.

a) Observed distribution map for Arenaviridae

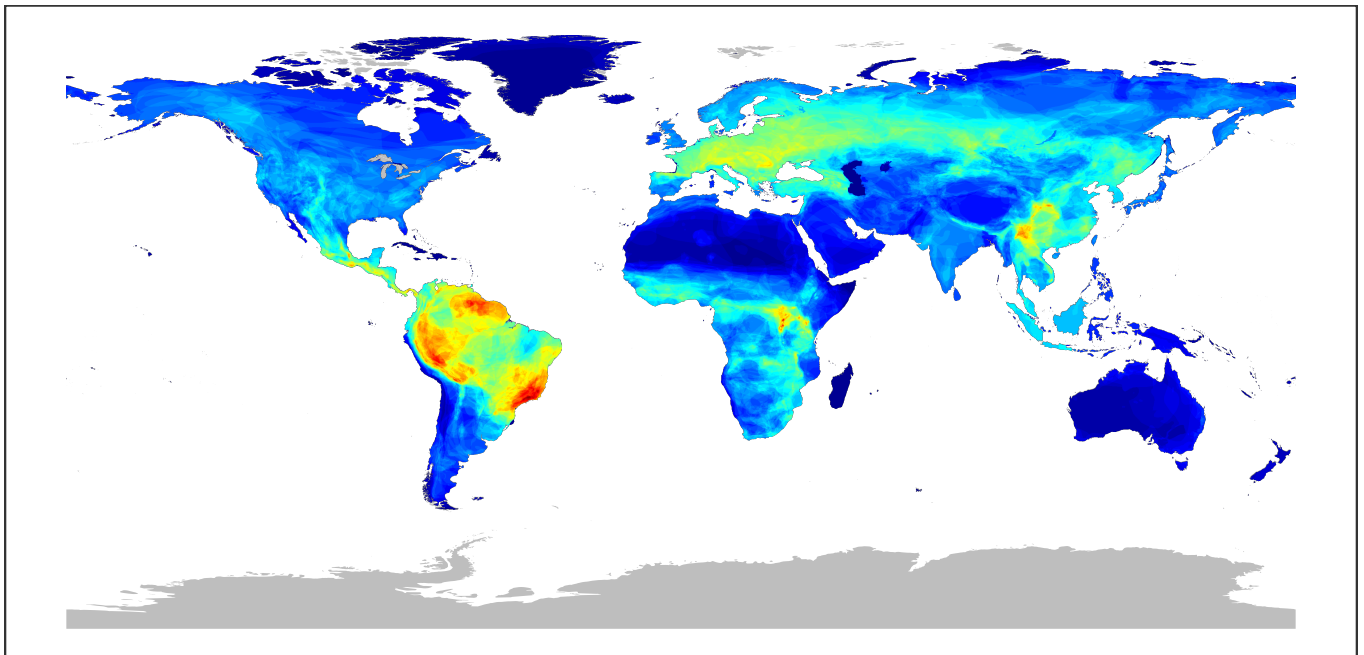


Number of observed susceptible species




0 10 20 30 40 50

b) Predicted distribution map for Arenaviridae ($p > 0.5$)



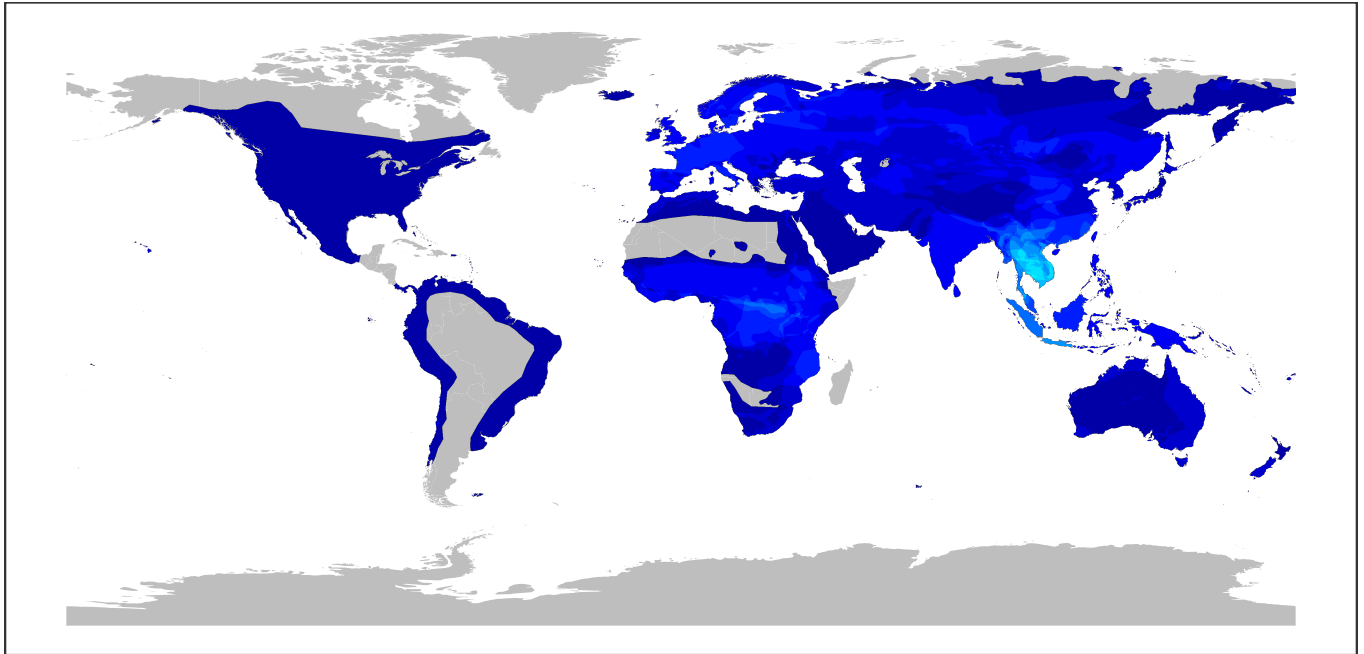
Number of observed and predicted susceptible species




0 10 20 30 40 50

Supplementary Figure SR62 | Geographic distribution of associations for Arenaviridae. The panels show the observed (a) and predicted (b) richness of mammal species associated with the viral family. A mammal species was considered predicted to be associated with the viral family if the median predicted probability exceeded 0.5.

a) Observed distribution map for Arteriviridae

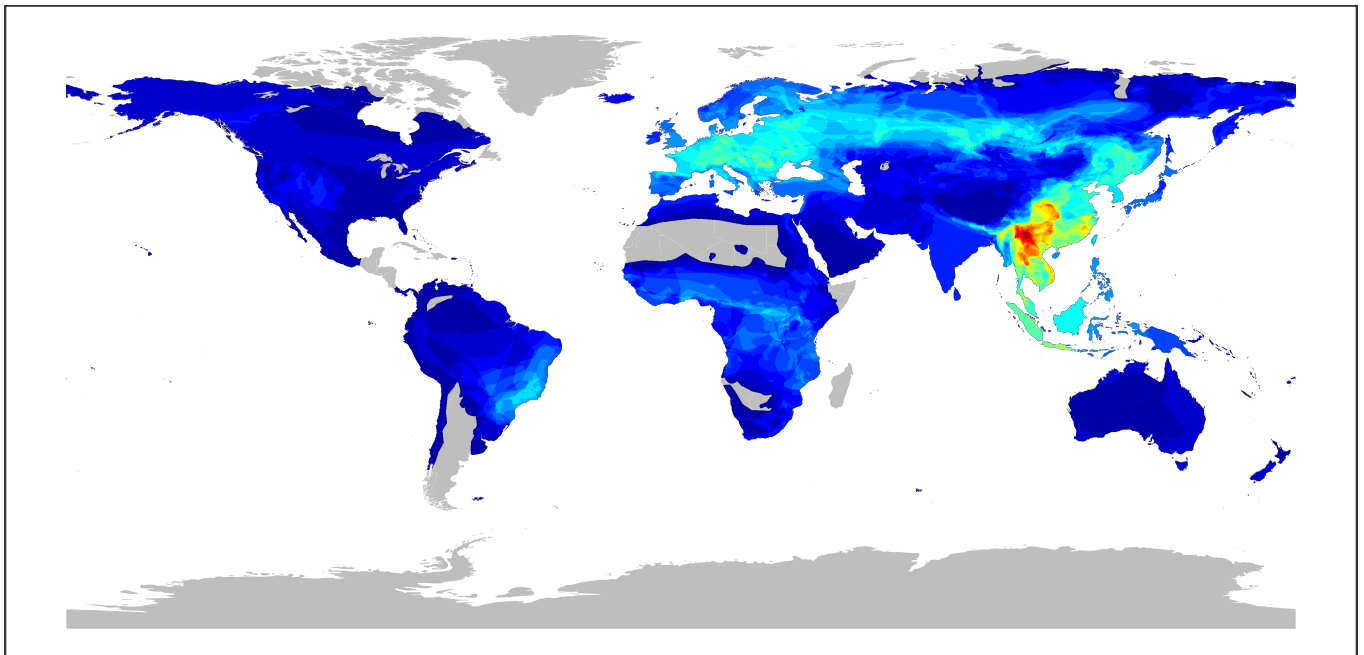


Number of observed susceptible species




0 5 10 15 20 25

b) Predicted distribution map for Arteriviridae ($p > 0.5$)



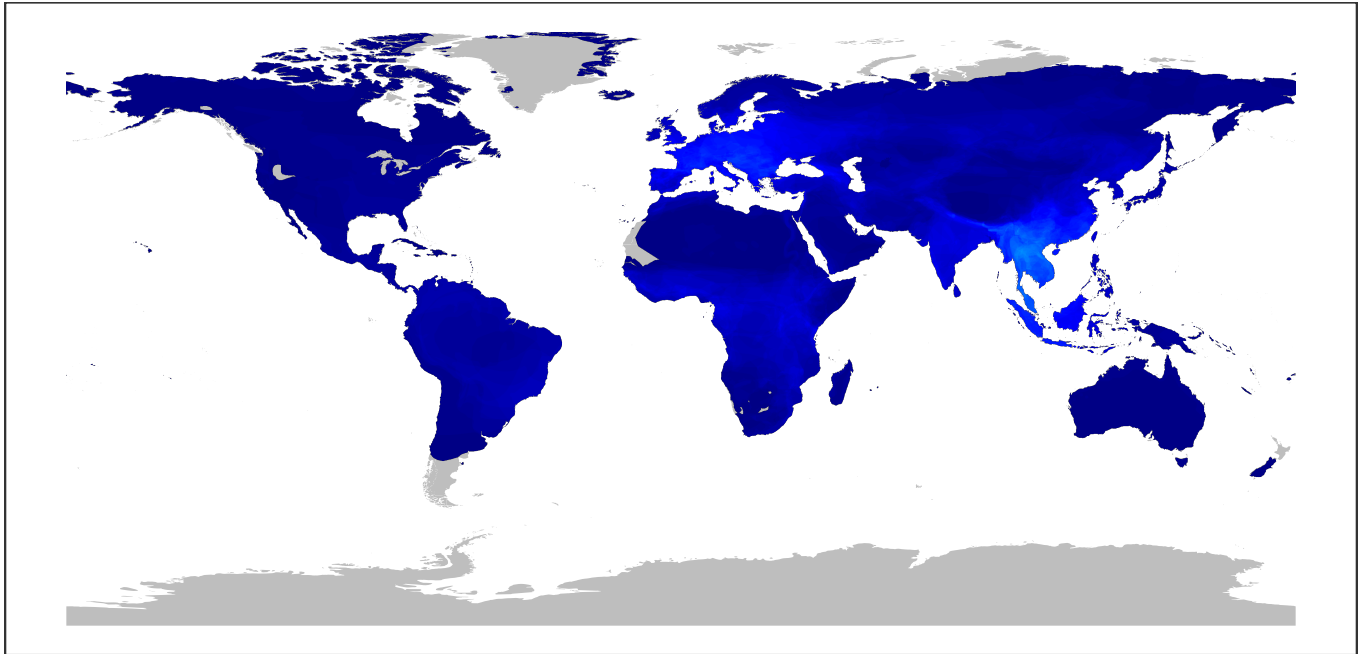
Number of observed and predicted susceptible species




0 5 10 15 20 25

Supplementary Figure SR63 | Geographic distribution of associations for Arteriviridae. The panels show the observed (a) and predicted (b) richness of mammal species associated with the viral family. A mammal species was considered predicted to be associated with the viral family if the median predicted probability exceeded 0.5.

a) Observed distribution map for Astroviridae

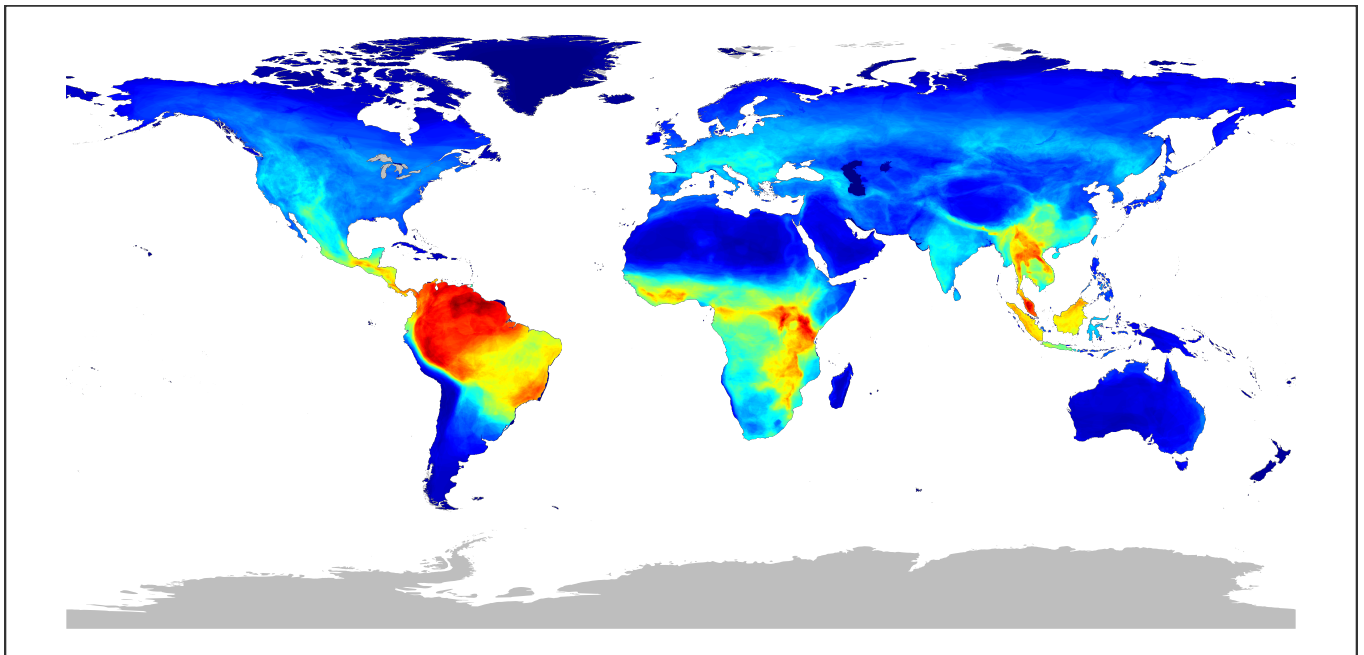


Number of observed susceptible species




0 50 100 150

b) Predicted distribution map for Astroviridae ($p > 0.5$)



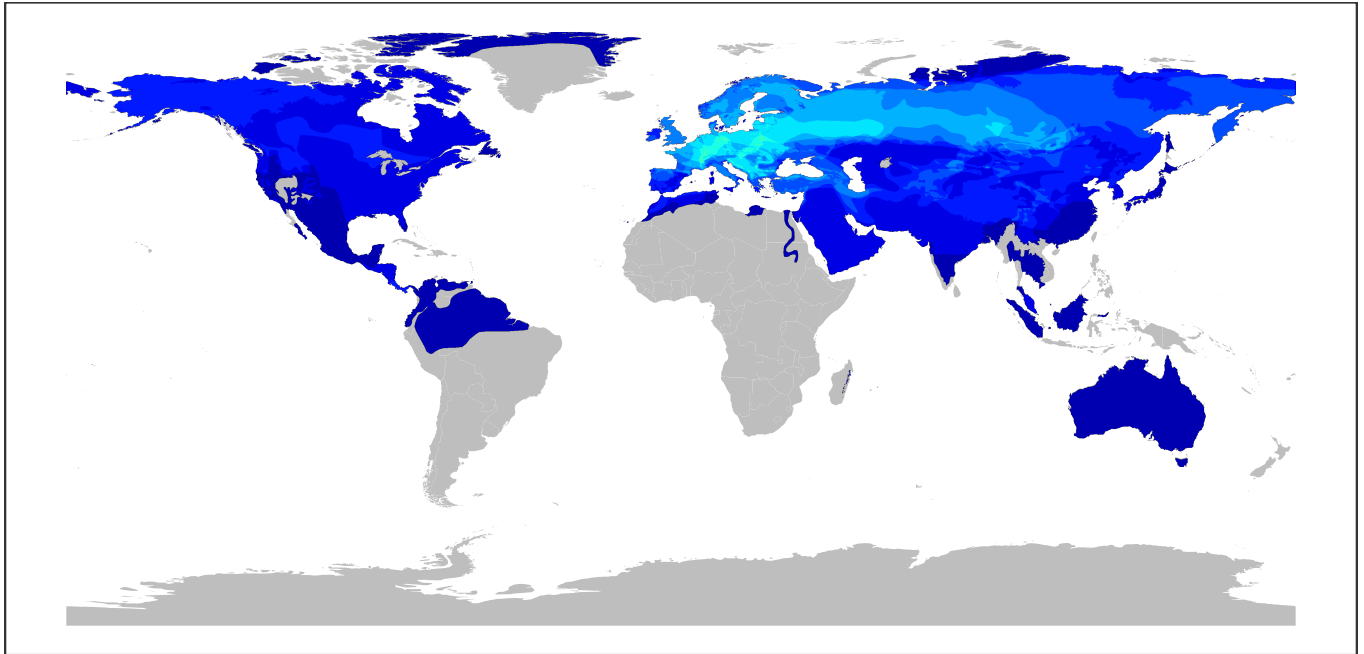
Number of observed and predicted susceptible species




0 50 100 150

Supplementary Figure SR64 | Geographic distribution of associations for Astroviridae. The panels show the observed (a) and predicted (b) richness of mammal species associated with the viral family. A mammal species was considered predicted to be associated with the viral family if the median predicted probability exceeded 0.5.

a) Observed distribution map for Bornaviridae

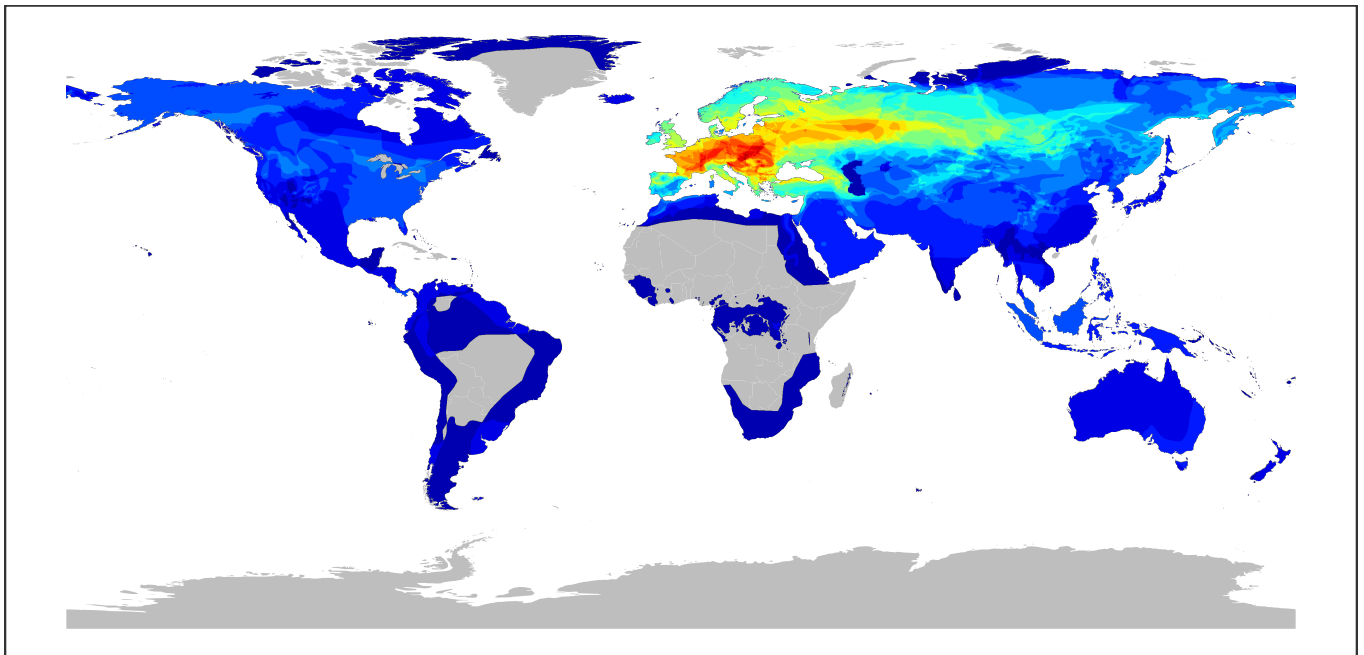


Number of observed susceptible species




0 5 10 15 20

b) Predicted distribution map for Bornaviridae ($p > 0.5$)



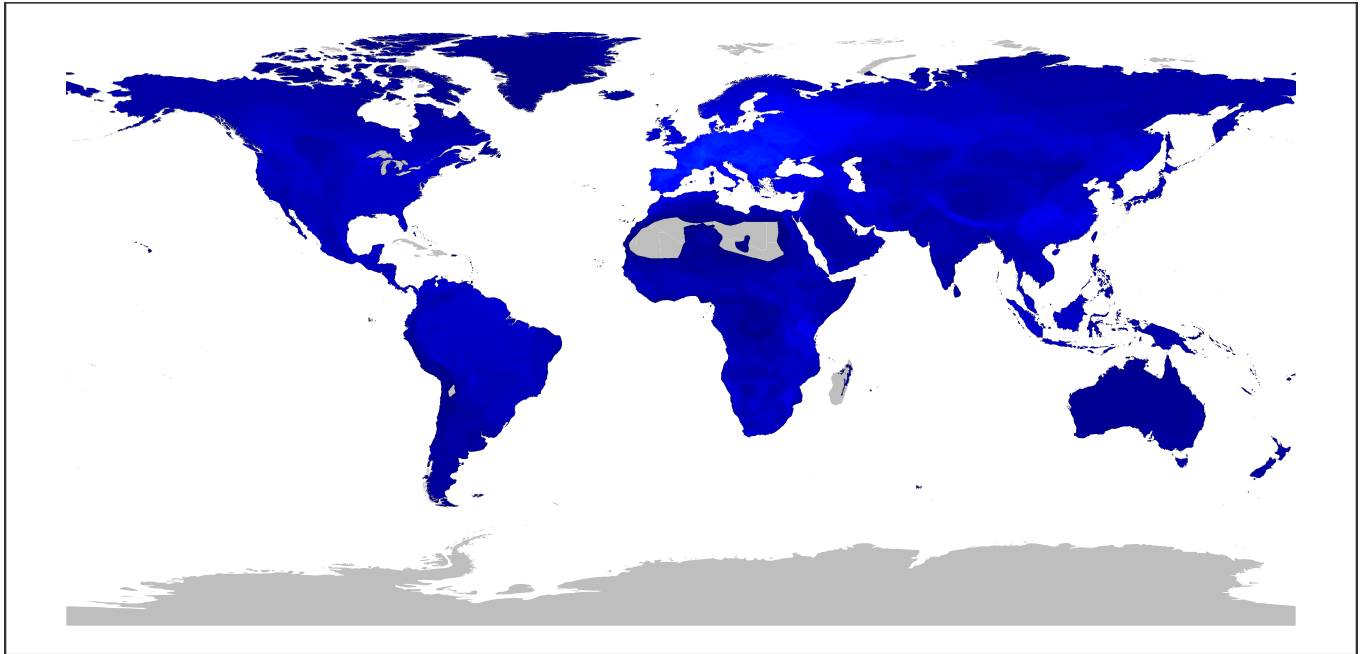
Number of observed and predicted susceptible species




0 5 10 15 20

Supplementary Figure SR65 | Geographic distribution of associations for Bornaviridae. The panels show the observed (**a**) and predicted (**b**) richness of mammal species associated with the viral family. A mammal species was considered predicted to be associated with the viral family if the median predicted probability exceeded 0.5.

a) Observed distribution map for Caliciviridae

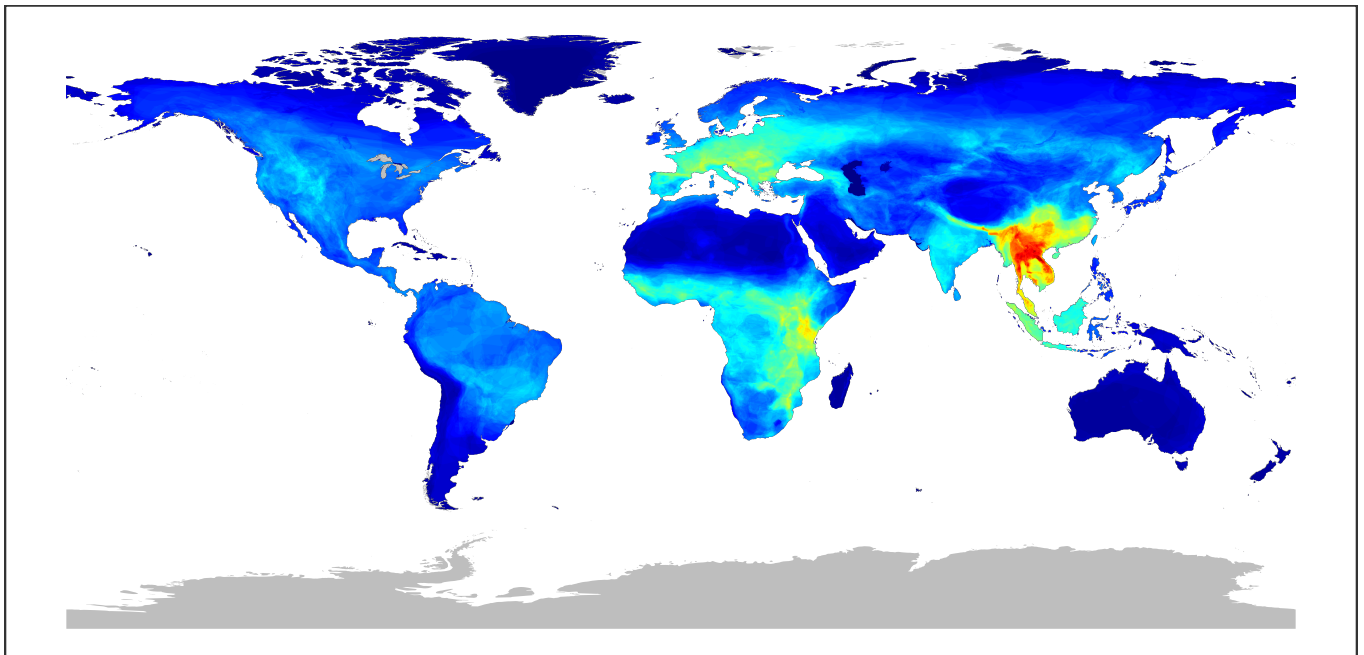


Number of observed susceptible species




0 25 50 75 100

b) Predicted distribution map for Caliciviridae ($p > 0.5$)



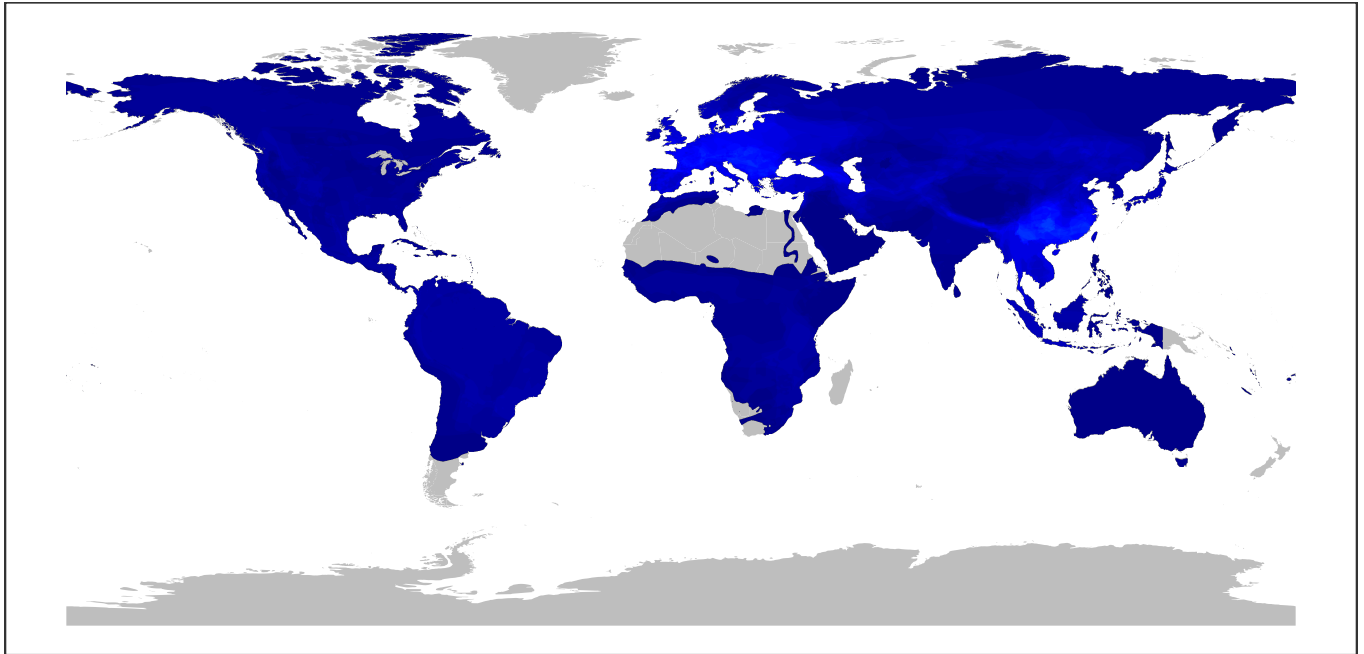
Number of observed and predicted susceptible species




0 25 50 75 100

Supplementary Figure SR66 | Geographic distribution of associations for Caliciviridae. The panels show the observed (a) and predicted (b) richness of mammal species associated with the viral family. A mammal species was considered predicted to be associated with the viral family if the median predicted probability exceeded 0.5.

a) Observed distribution map for Circoviridae

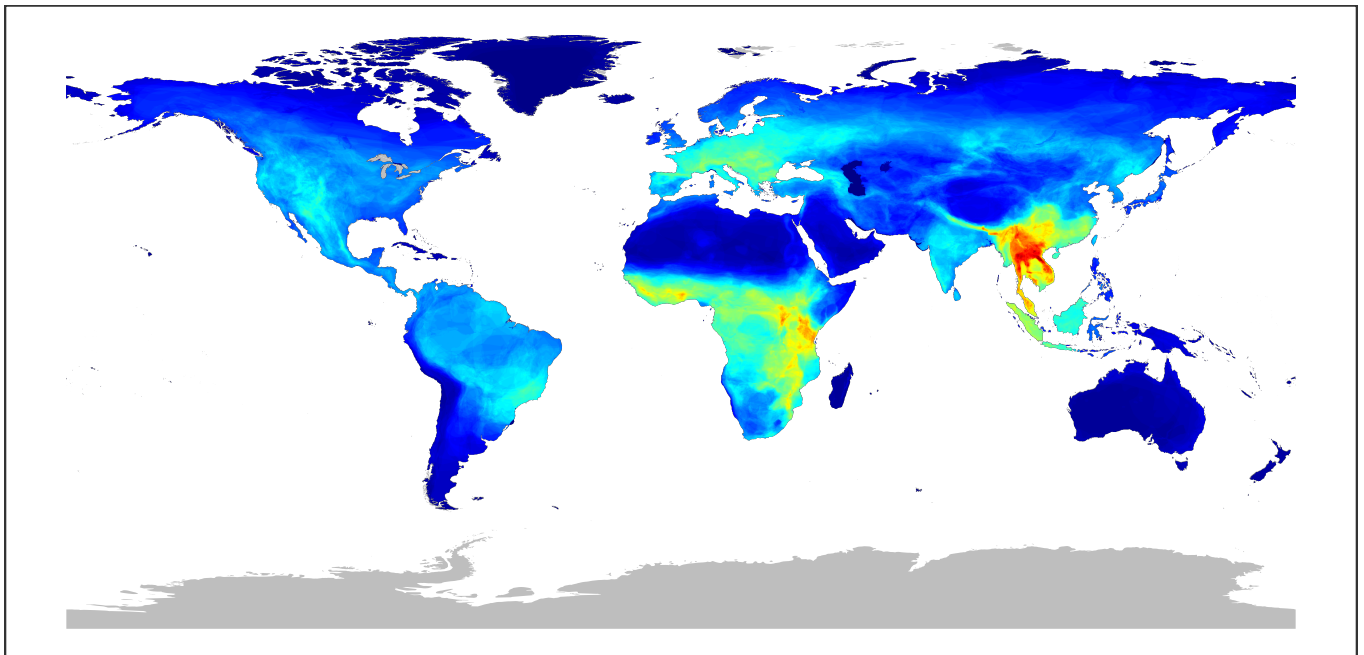


Number of observed susceptible species




0 30 60 90 120

b) Predicted distribution map for Circoviridae ($p > 0.5$)



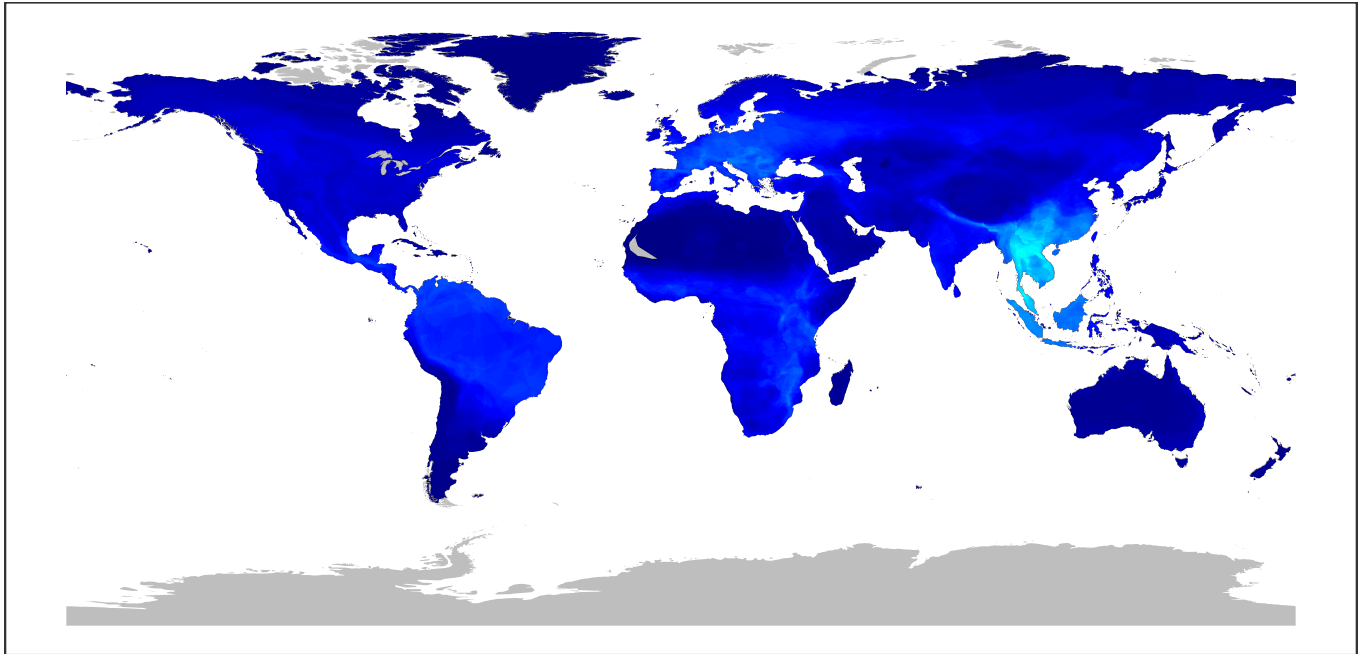
Number of observed and predicted susceptible species




0 30 60 90 120

Supplementary Figure SR67 | Geographic distribution of associations for Circoviridae. The panels show the observed (a) and predicted (b) richness of mammal species associated with the viral family. A mammal species was considered predicted to be associated with the viral family if the median predicted probability exceeded 0.5.

a) Observed distribution map for Coronaviridae

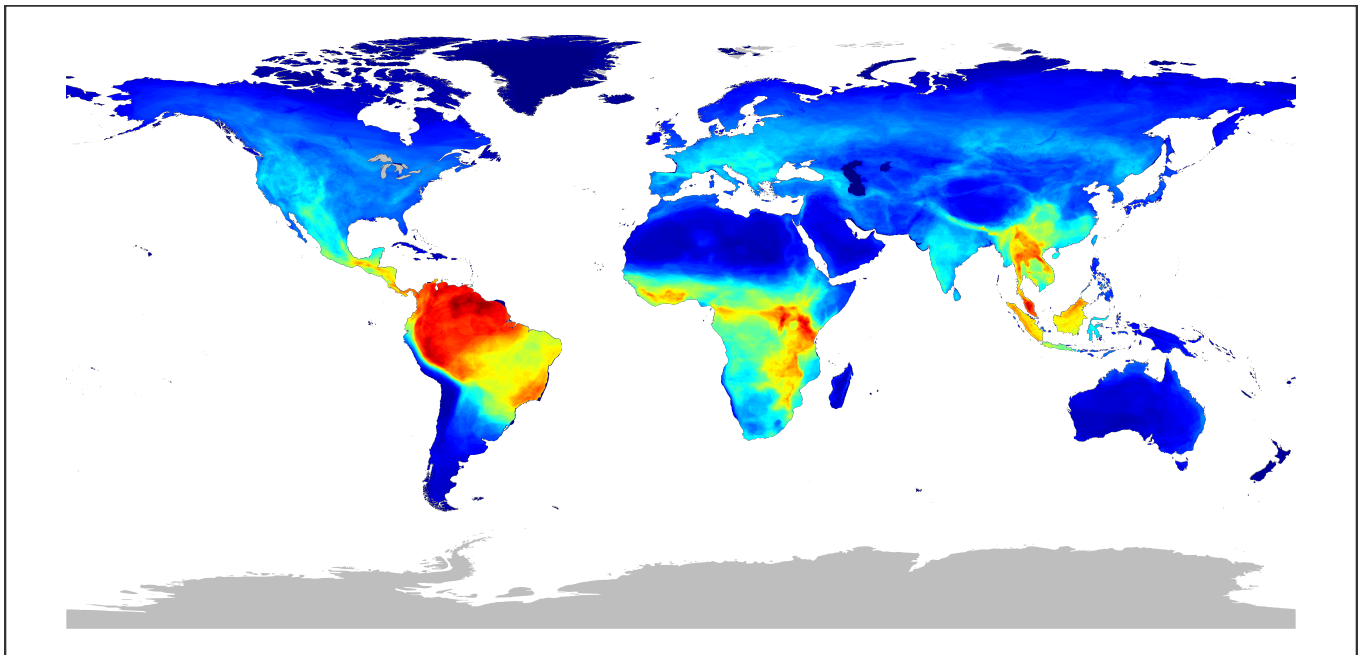


Number of observed susceptible species




0 40 80 120

b) Predicted distribution map for Coronaviridae ($p > 0.5$)



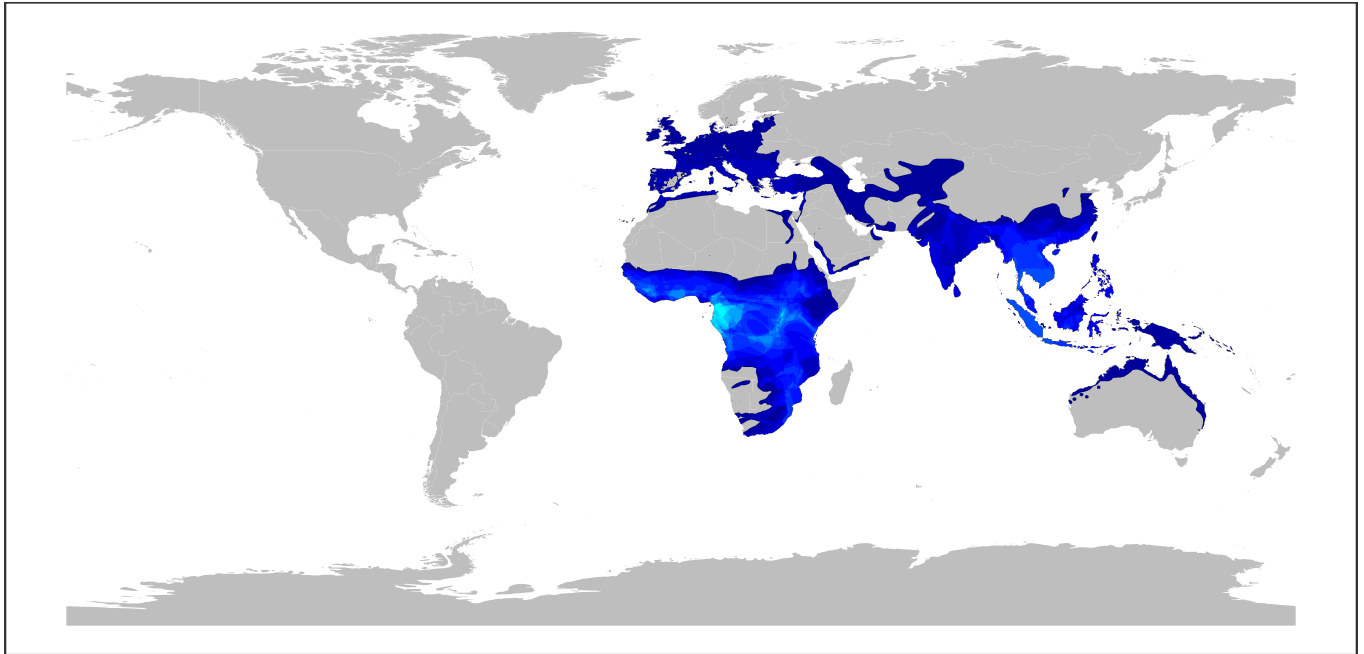
Number of observed and predicted susceptible species




0 40 80 120

Supplementary Figure SR68 | Geographic distribution of associations for Coronaviridae. The panels show the observed (**a**) and predicted (**b**) richness of mammal species associated with the viral family. A mammal species was considered predicted to be associated with the viral family if the median predicted probability exceeded 0.5.

a) Observed distribution map for Filoviridae

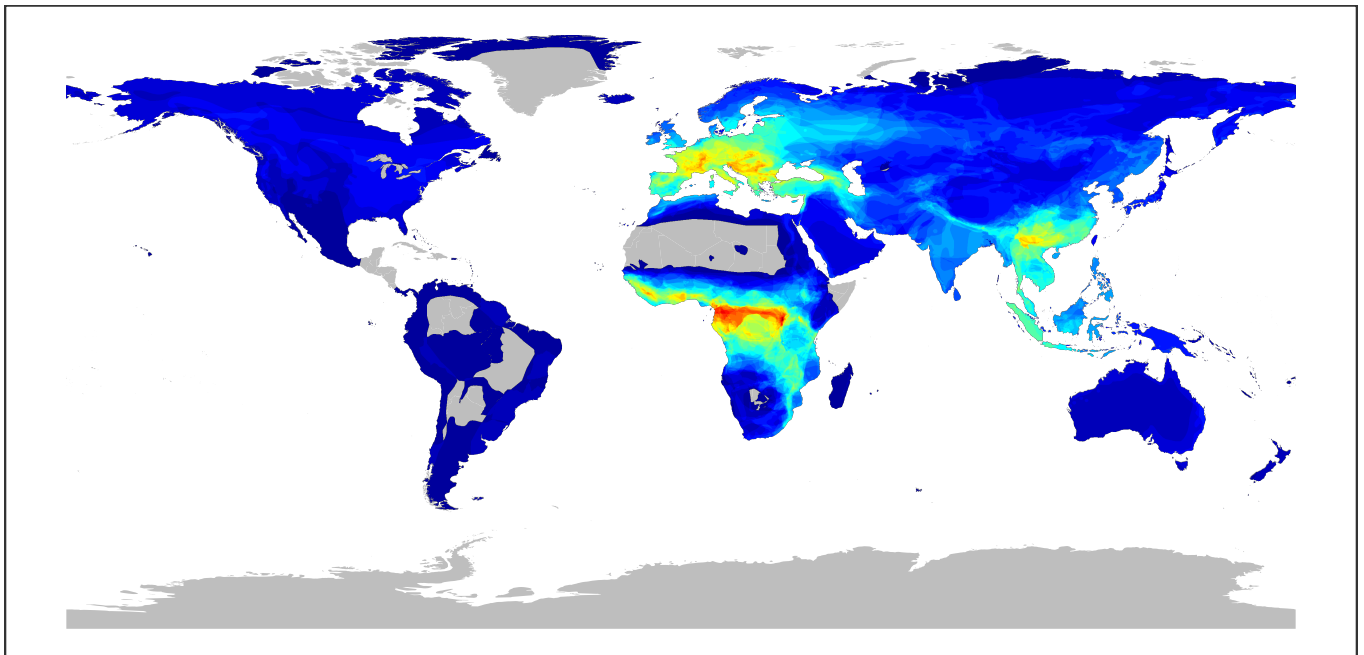


Number of observed susceptible species




0 10 20 30

b) Predicted distribution map for Filoviridae ($p > 0.5$)



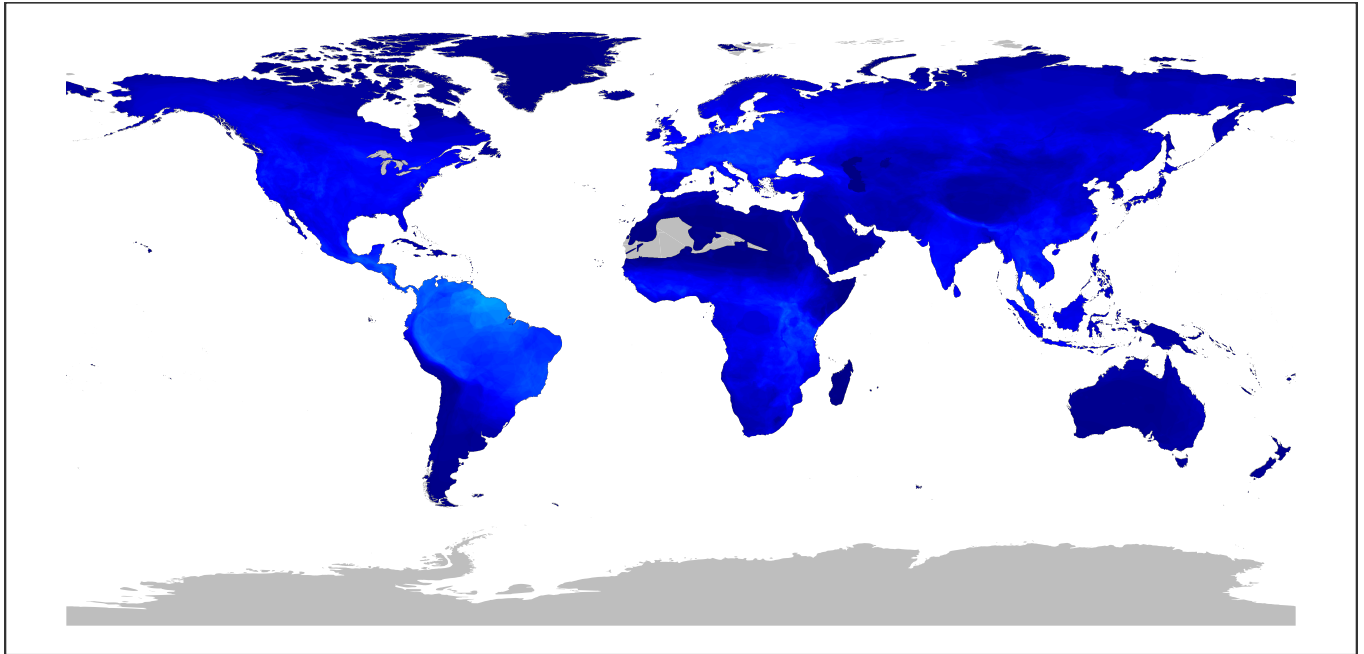
Number of observed and predicted susceptible species




0 10 20 30

Supplementary Figure SR69 | Geographic distribution of associations for Filoviridae. The panels show the observed (**a**) and predicted (**b**) richness of mammal species associated with the viral family. A mammal species was considered predicted to be associated with the viral family if the median predicted probability exceeded 0.5.

a) Observed distribution map for Flaviviridae

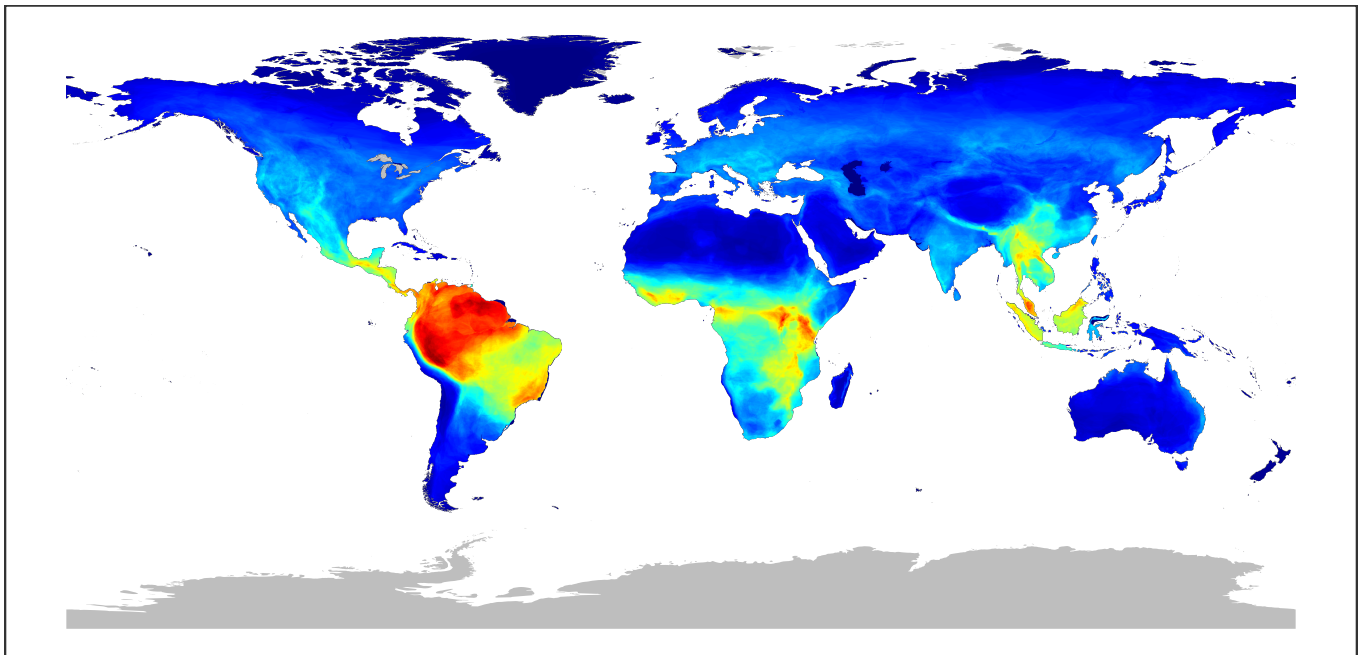


Number of observed susceptible species




0 50 100 150

b) Predicted distribution map for Flaviviridae ($p > 0.5$)



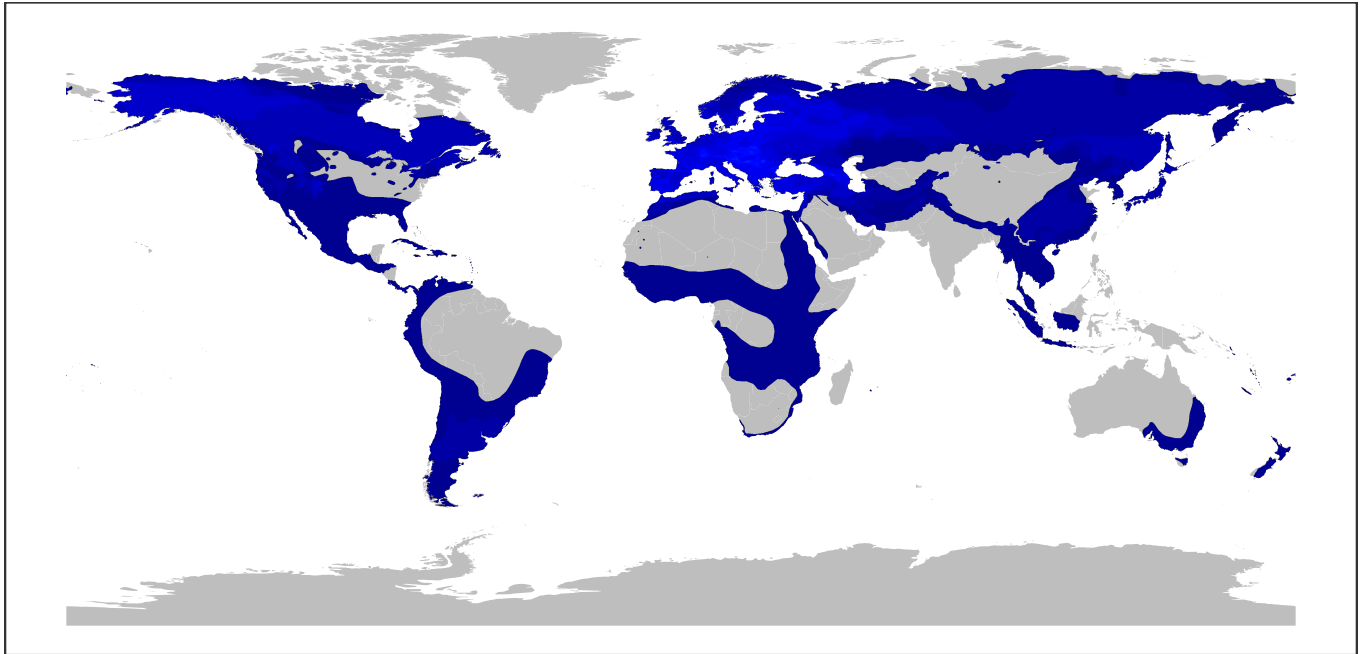
Number of observed and predicted susceptible species




0 50 100 150

Supplementary Figure SR70 | Geographic distribution of associations for Flaviviridae. The panels show the observed (a) and predicted (b) richness of mammal species associated with the viral family. A mammal species was considered predicted to be associated with the viral family if the median predicted probability exceeded 0.5.

a) Observed distribution map for Genomoviridae

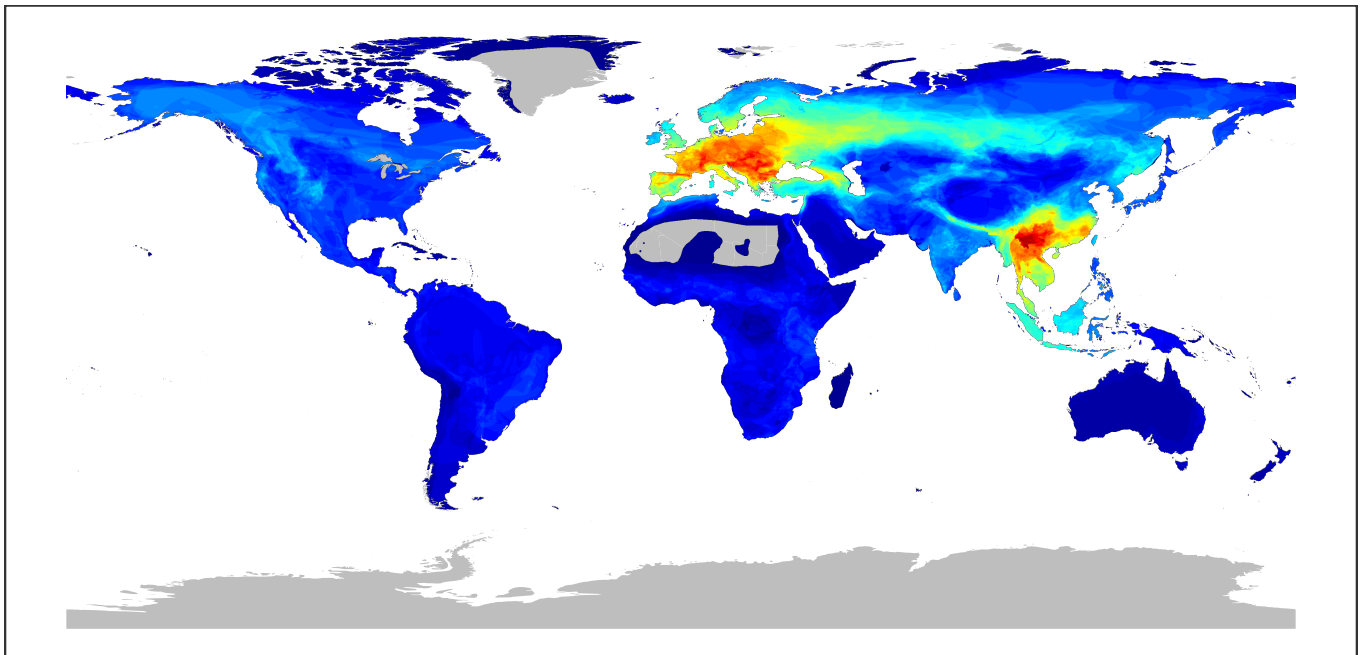


Number of observed susceptible species




0 10 20 30 40 50

b) Predicted distribution map for Genomoviridae ($p > 0.5$)



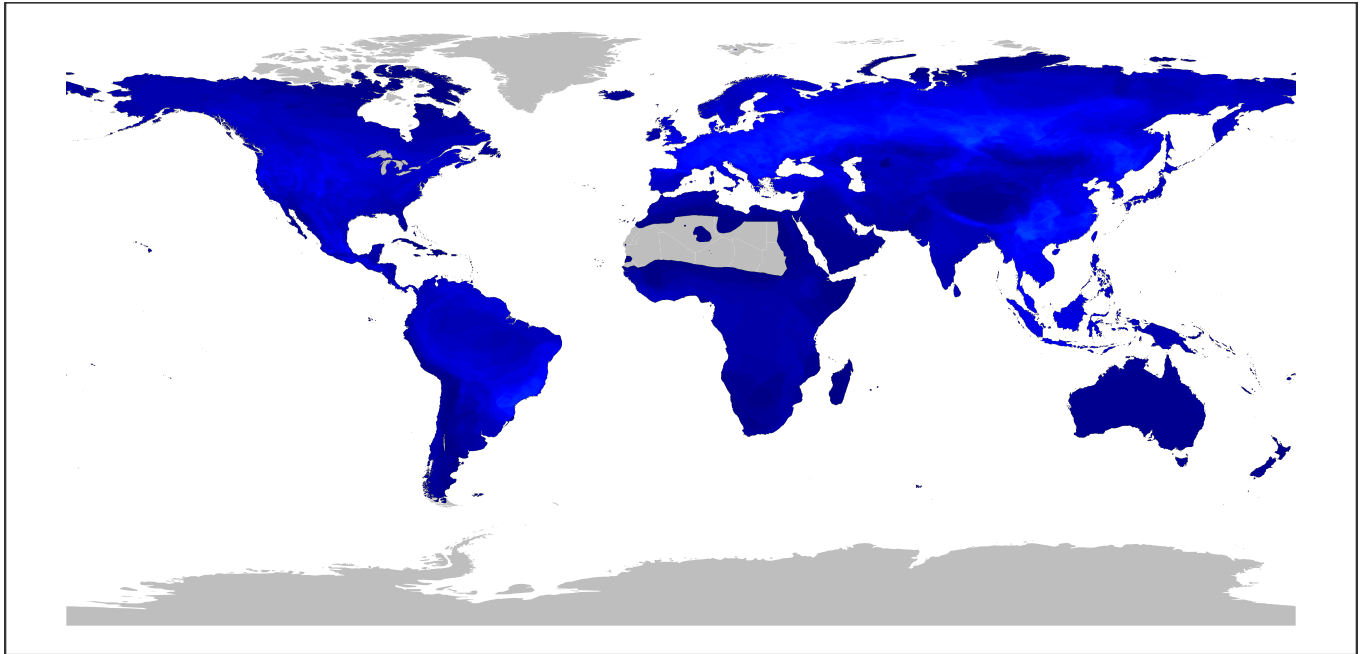
Number of observed and predicted susceptible species




0 10 20 30 40 50

Supplementary Figure SR71 | Geographic distribution of associations for Genomoviridae. The panels show the observed (**a**) and predicted (**b**) richness of mammal species associated with the viral family. A mammal species was considered predicted to be associated with the viral family if the median predicted probability exceeded 0.5.

a) Observed distribution map for Hantaviridae

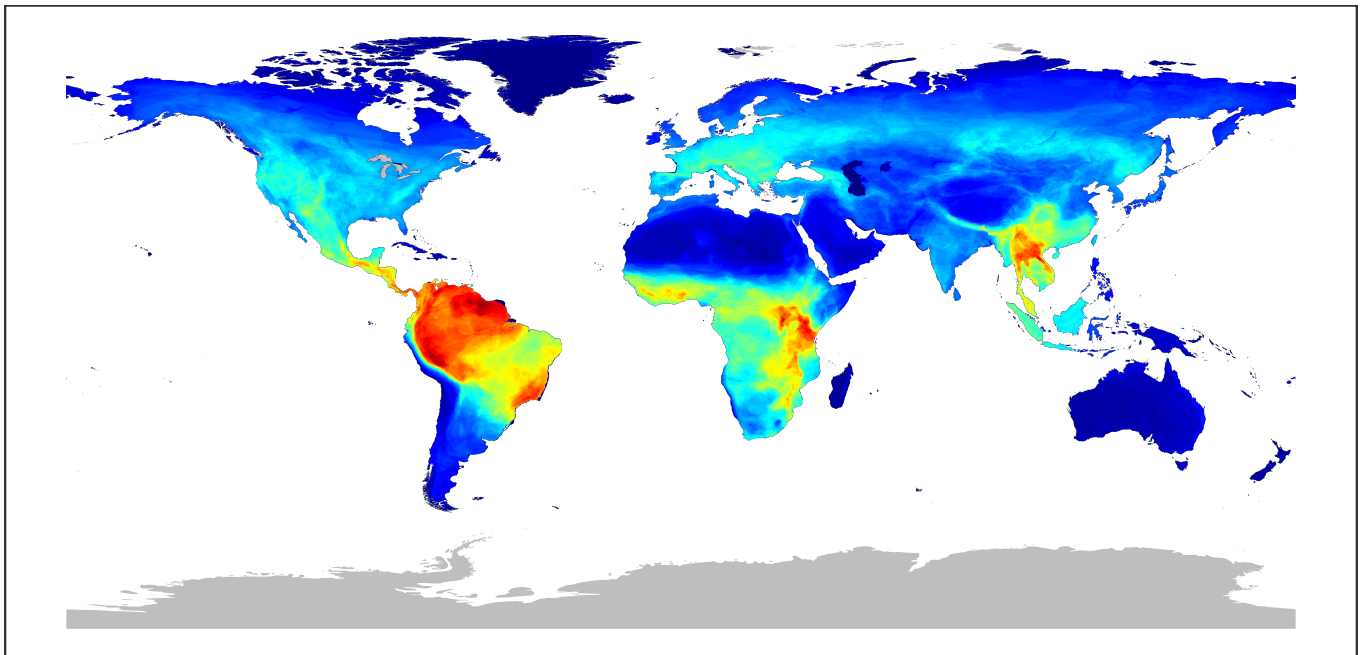


Number of observed susceptible species




0 50 100

b) Predicted distribution map for Hantaviridae ($p > 0.5$)



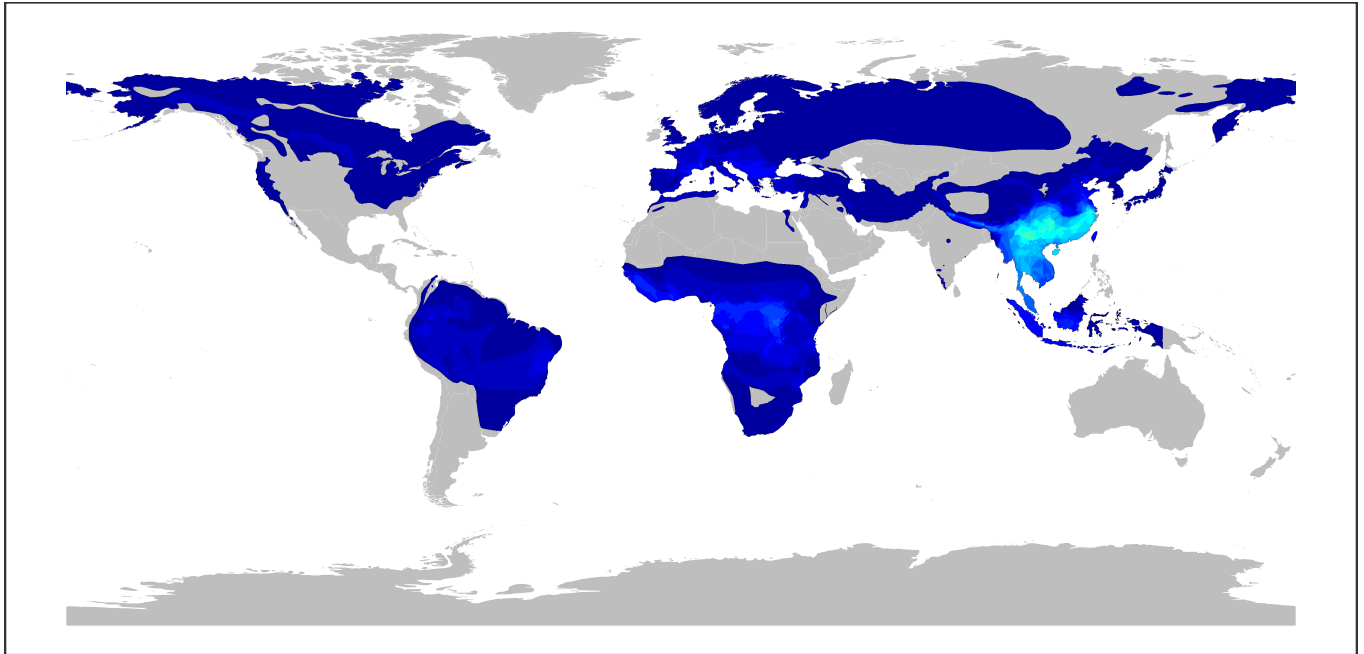
Number of observed and predicted susceptible species




0 50 100

Supplementary Figure SR72 | Geographic distribution of associations for Hantaviridae. The panels show the observed (a) and predicted (b) richness of mammal species associated with the viral family. A mammal species was considered predicted to be associated with the viral family if the median predicted probability exceeded 0.5.

a) Observed distribution map for Hepadnaviridae

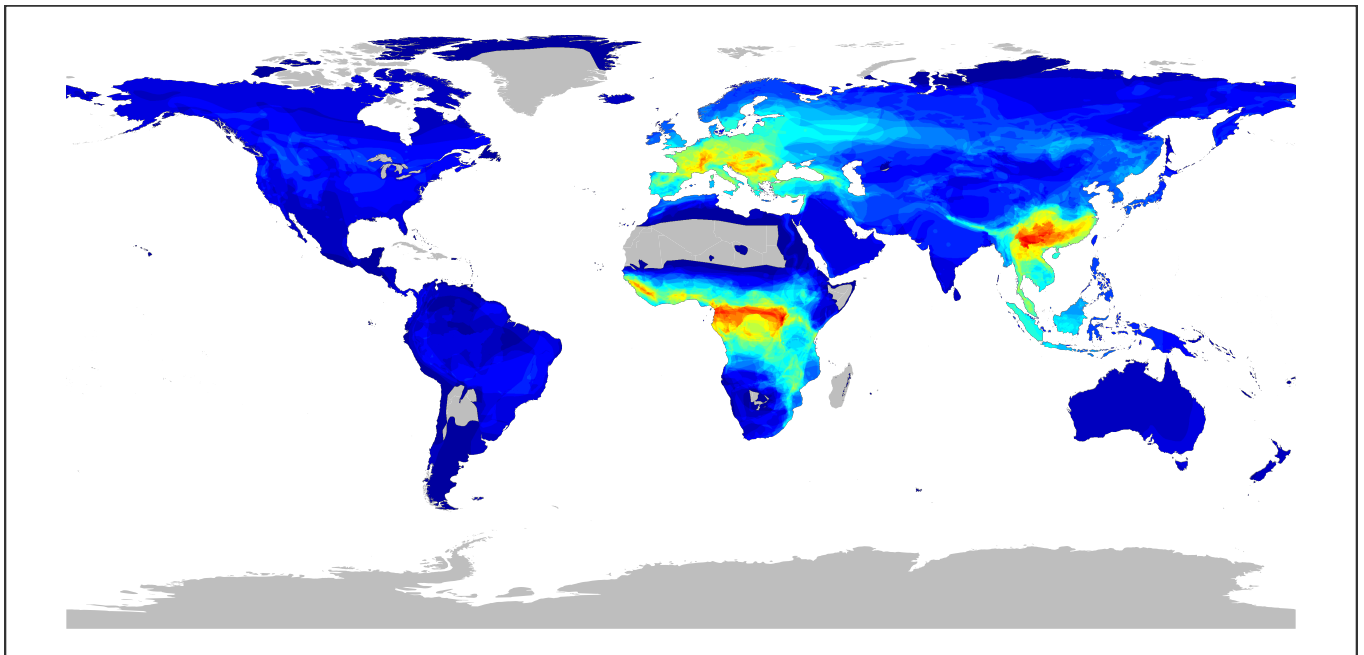


Number of observed susceptible species




0 10 20 30

b) Predicted distribution map for Hepadnaviridae ($p > 0.5$)



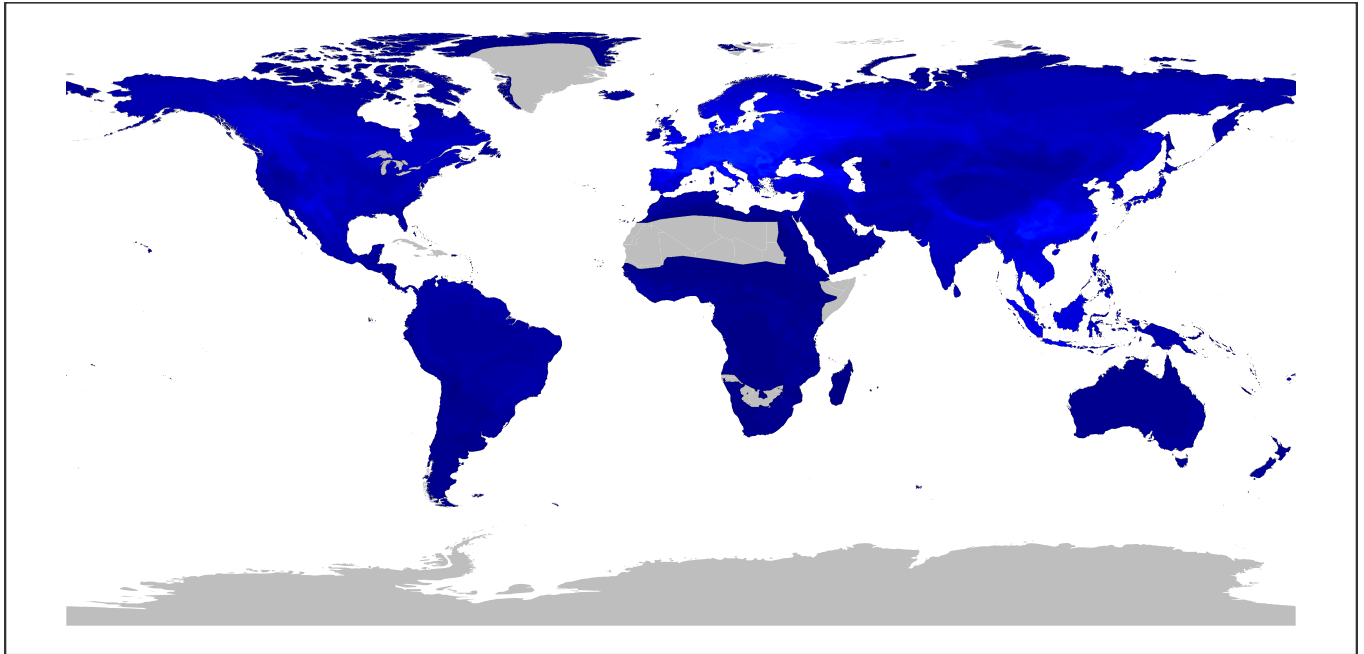
Number of observed and predicted susceptible species




0 10 20 30

Supplementary Figure SR73 | Geographic distribution of associations for Hepadnaviridae. The panels show the observed (**a**) and predicted (**b**) richness of mammal species associated with the viral family. A mammal species was considered predicted to be associated with the viral family if the median predicted probability exceeded 0.5.

a) Observed distribution map for Hepeviridae

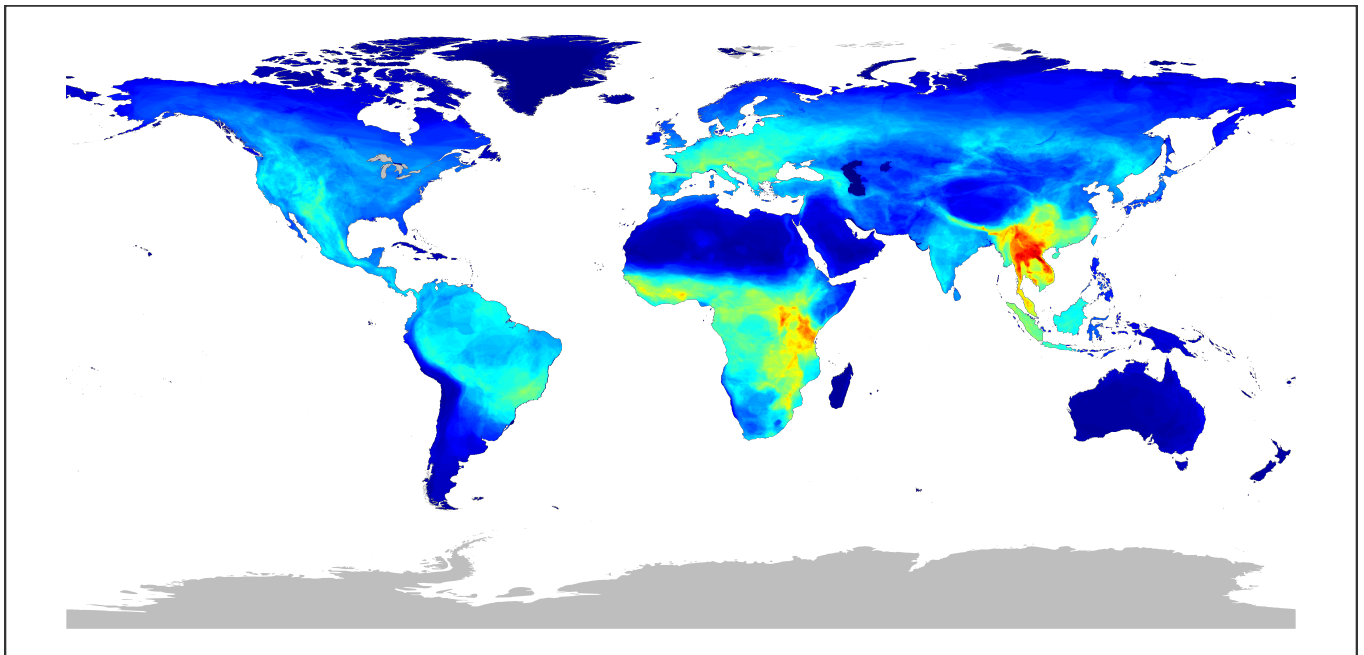


Number of observed susceptible species




0 30 60 90 120

b) Predicted distribution map for Hepeviridae ($p > 0.5$)



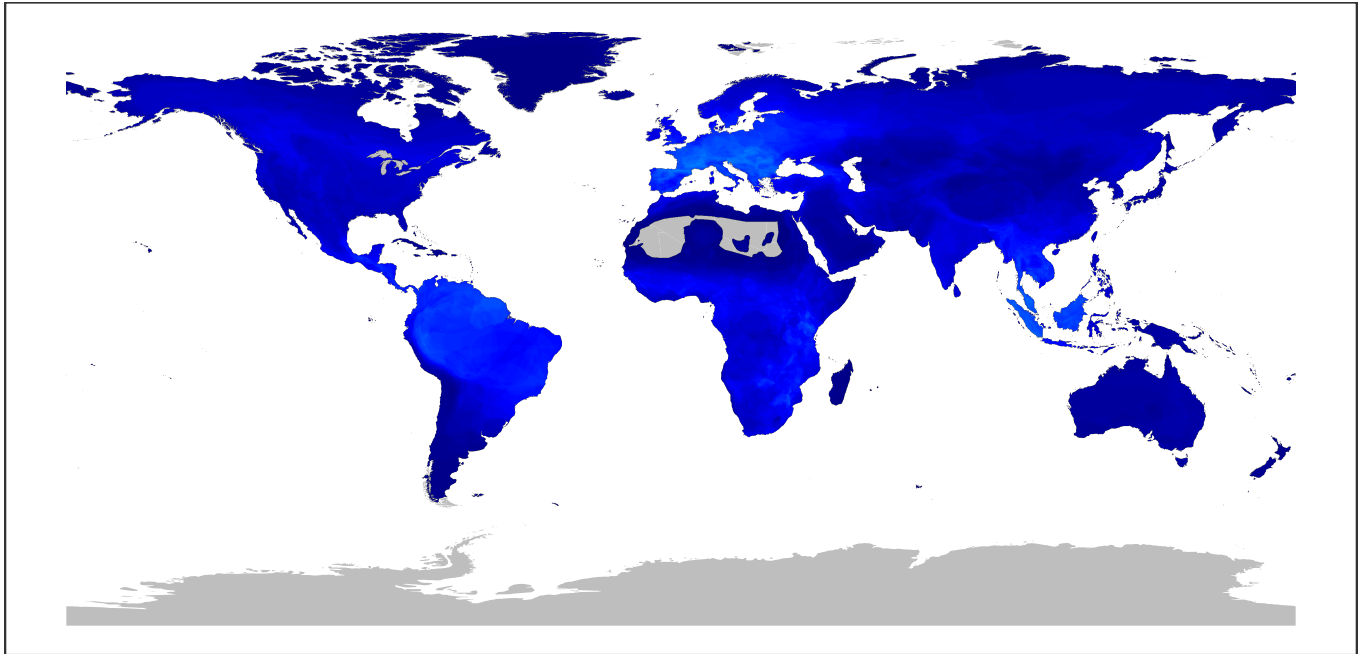
Number of observed and predicted susceptible species




0 30 60 90 120

Supplementary Figure SR74 | Geographic distribution of associations for Hepeviridae. The panels show the observed (a) and predicted (b) richness of mammal species associated with the viral family. A mammal species was considered predicted to be associated with the viral family if the median predicted probability exceeded 0.5.

a) Observed distribution map for Herpesviridae

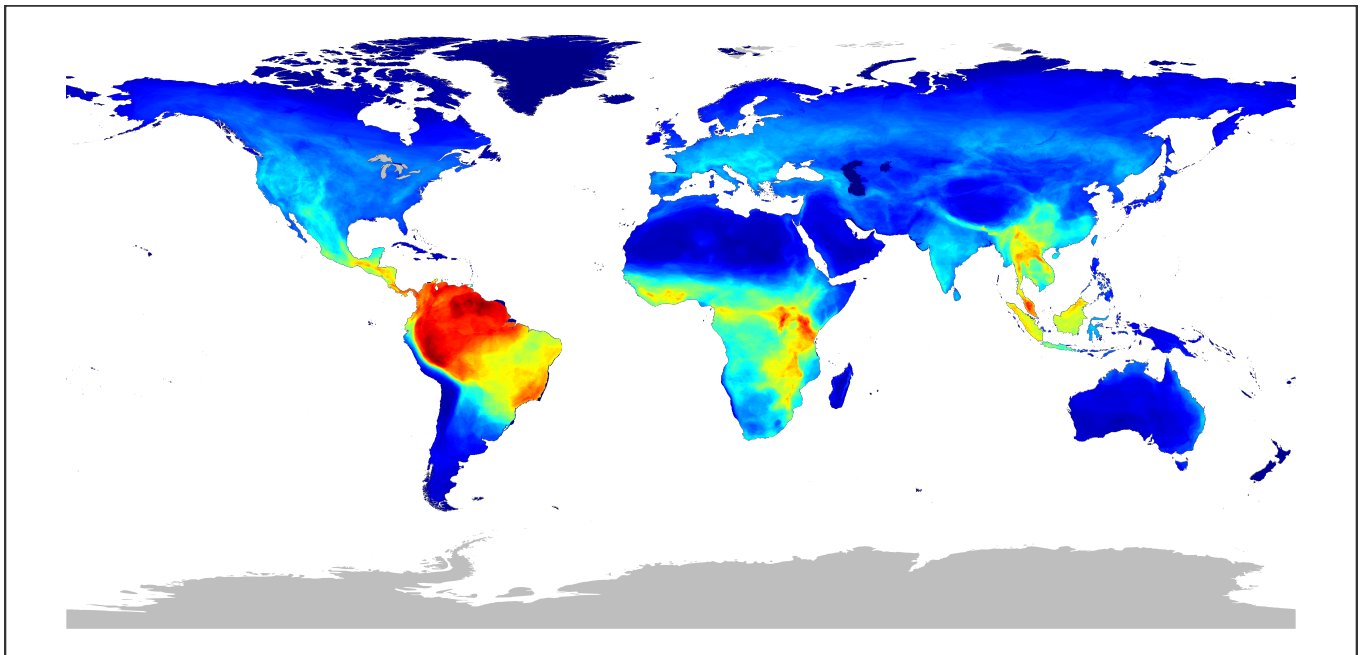


Number of observed susceptible species




0 50 100 150

b) Predicted distribution map for Herpesviridae ($p > 0.5$)



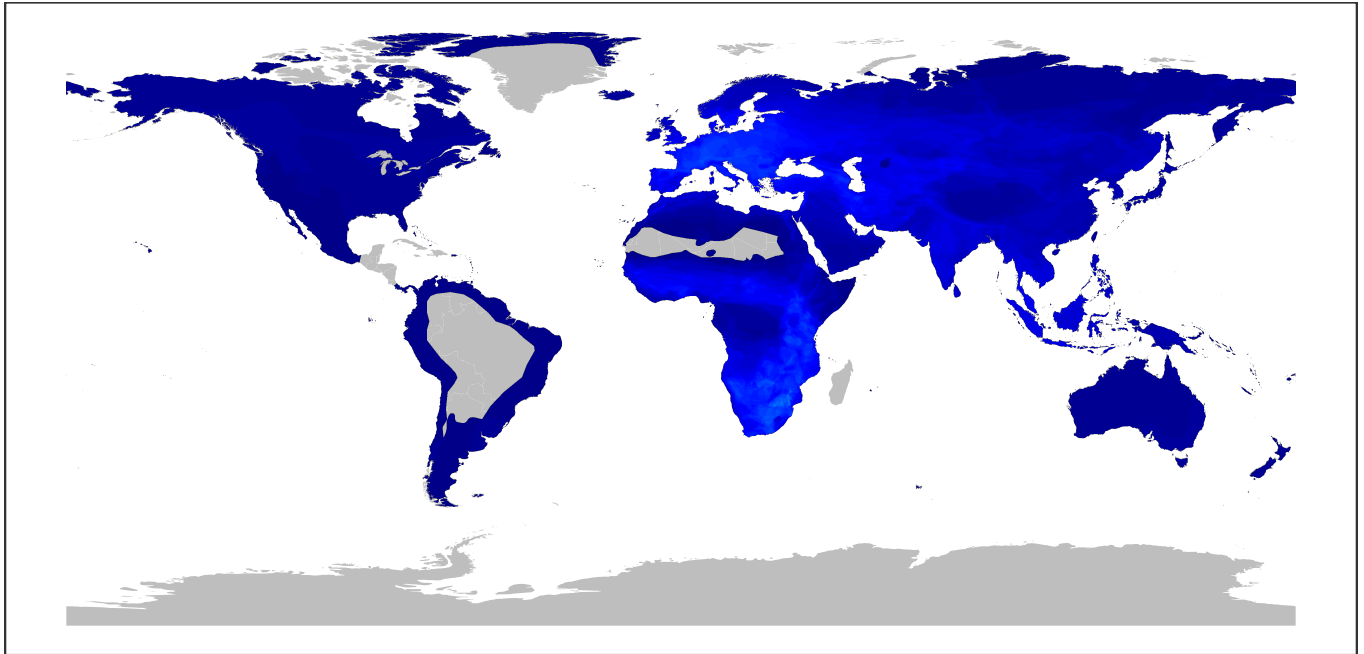
Number of observed and predicted susceptible species




0 50 100 150

Supplementary Figure SR75 | Geographic distribution of associations for Herpesviridae. The panels show the observed (a) and predicted (b) richness of mammal species associated with the viral family. A mammal species was considered predicted to be associated with the viral family if the median predicted probability exceeded 0.5.

a) Observed distribution map for Nairoviridae

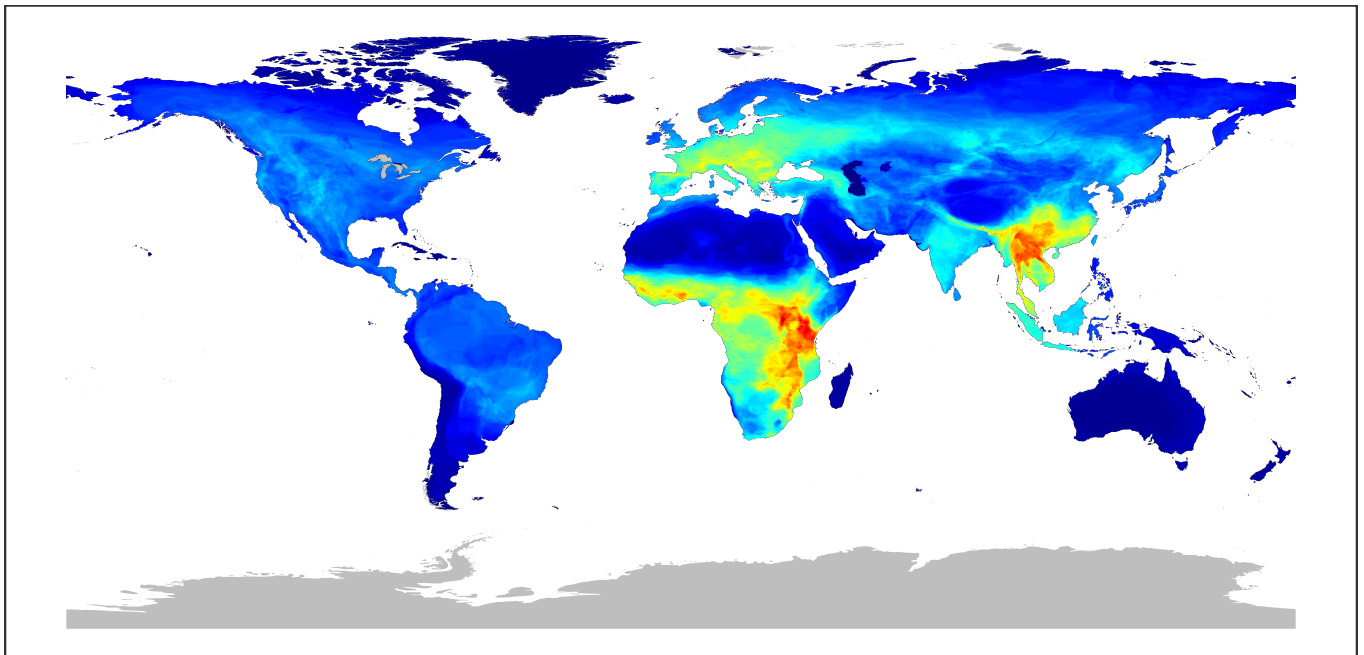


Number of observed susceptible species




0 25 50 75 100

b) Predicted distribution map for Nairoviridae ($p > 0.5$)



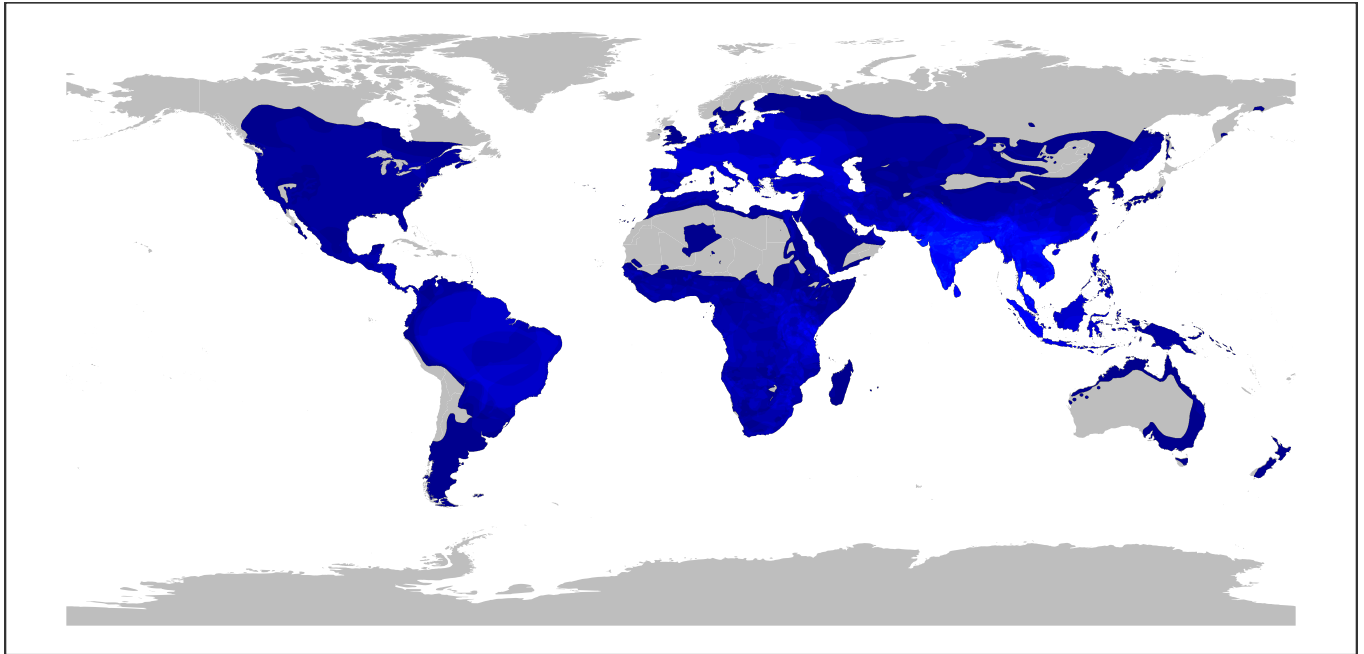
Number of observed and predicted susceptible species




0 25 50 75 100

Supplementary Figure SR76 | Geographic distribution of associations for Nairoviridae. The panels show the observed (**a**) and predicted (**b**) richness of mammal species associated with the viral family. A mammal species was considered predicted to be associated with the viral family if the median predicted probability exceeded 0.5.

a) Observed distribution map for Orthomyxoviridae

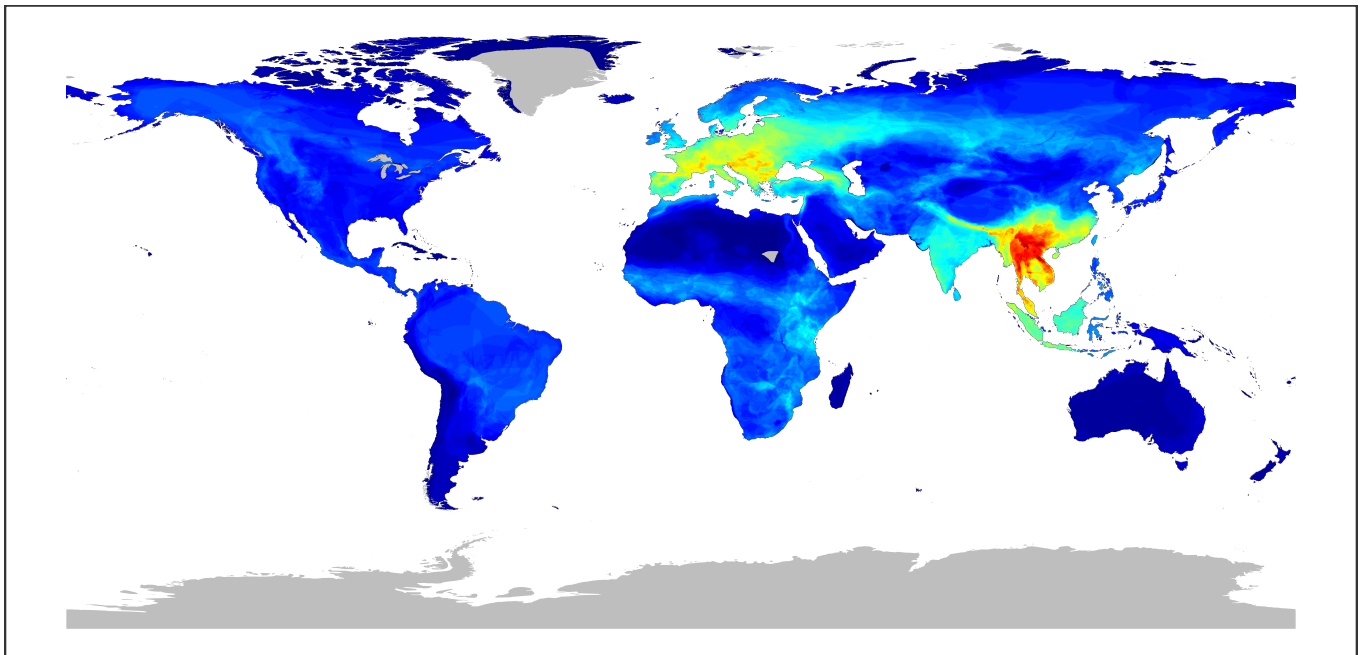


Number of observed susceptible species




0 20 40 60

b) Predicted distribution map for Orthomyxoviridae ($p > 0.5$)



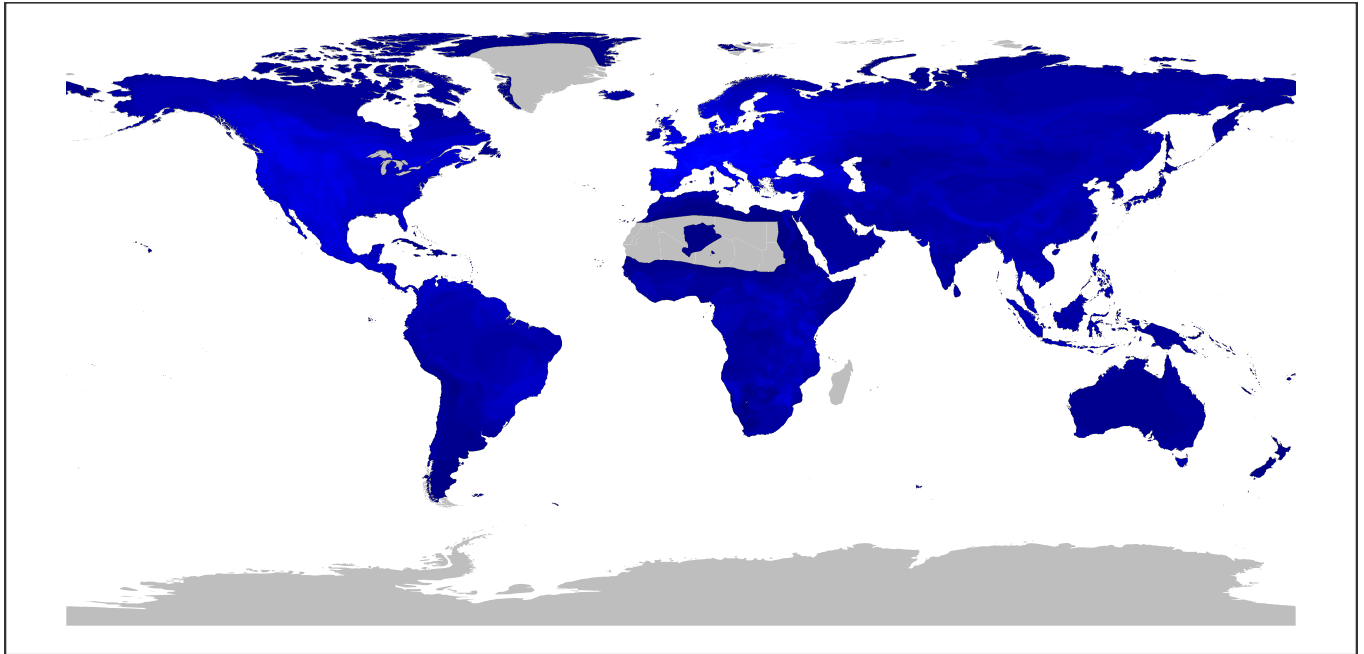
Number of observed and predicted susceptible species




0 20 40 60

Supplementary Figure SR77 | Geographic distribution of associations for Orthomyxoviridae. The panels show the observed (a) and predicted (b) richness of mammal species associated with the viral family. A mammal species was considered predicted to be associated with the viral family if the median predicted probability exceeded 0.5.

a) Observed distribution map for Papillomaviridae

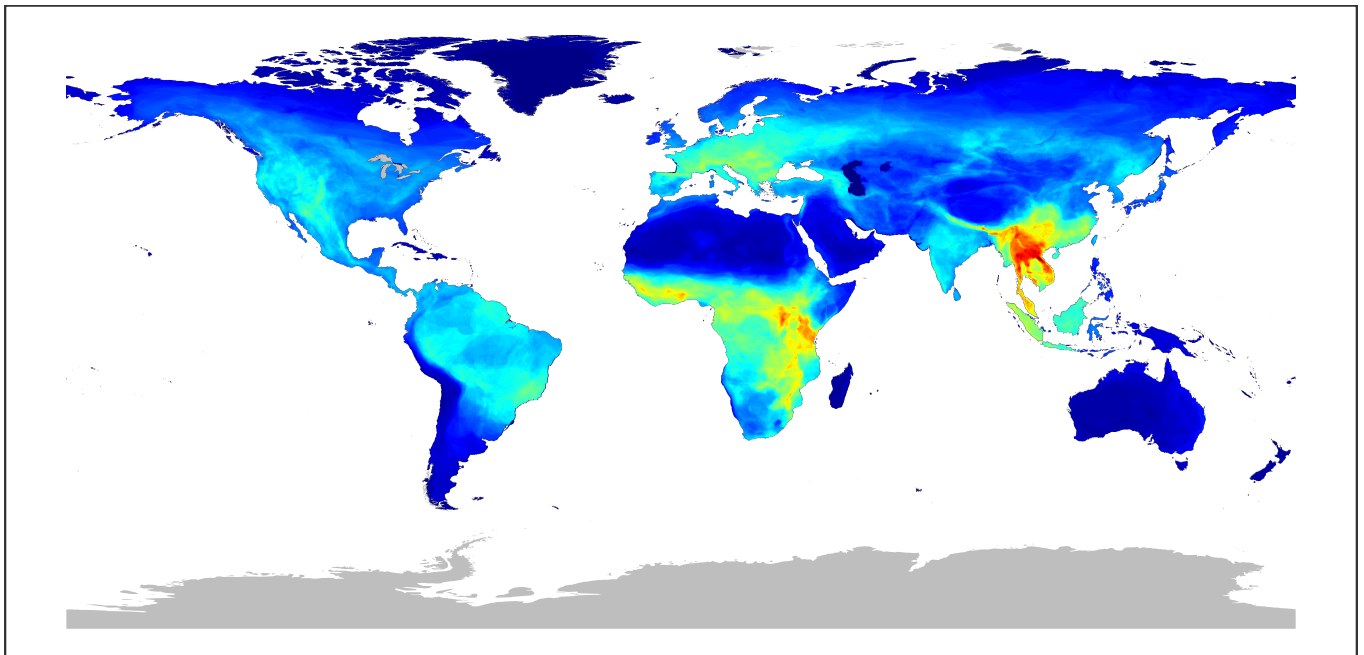


Number of observed susceptible species




0 30 60 90 120

b) Predicted distribution map for Papillomaviridae ($p > 0.5$)



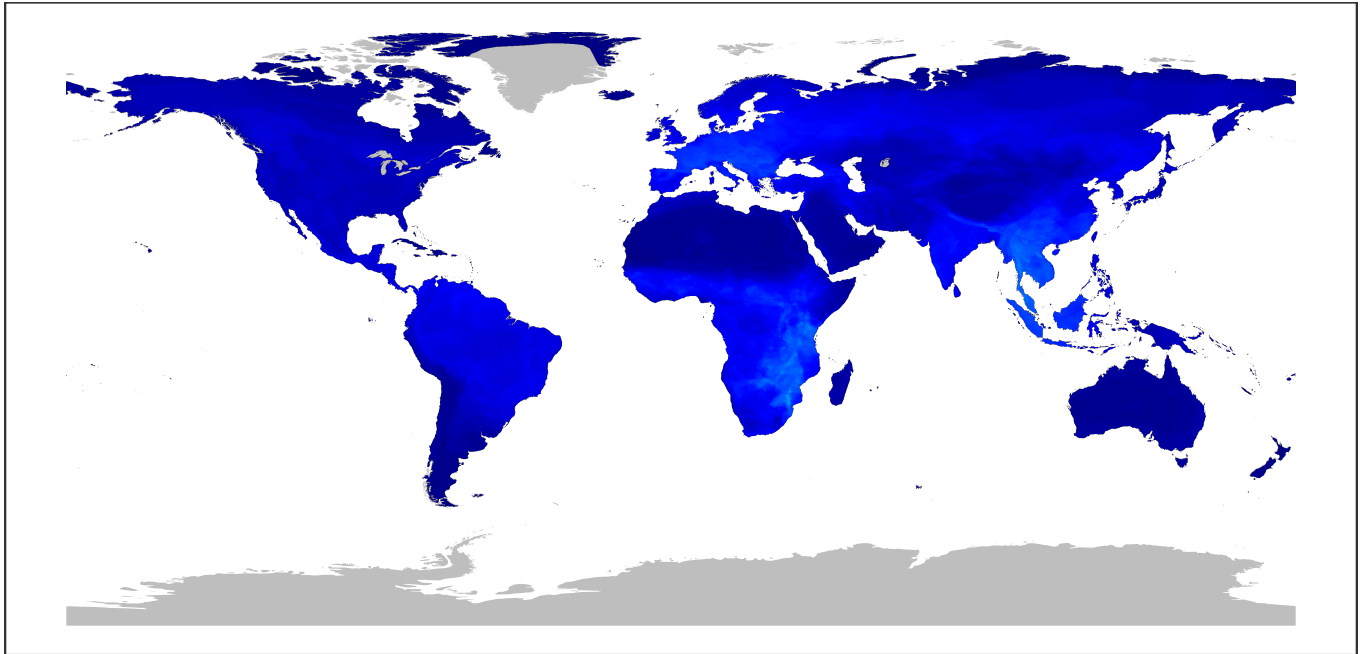
Number of observed and predicted susceptible species




0 30 60 90 120

Supplementary Figure SR78 | Geographic distribution of associations for Papillomaviridae. The panels show the observed (a) and predicted (b) richness of mammal species associated with the viral family. A mammal species was considered predicted to be associated with the viral family if the median predicted probability exceeded 0.5.

a) Observed distribution map for Paramyxoviridae

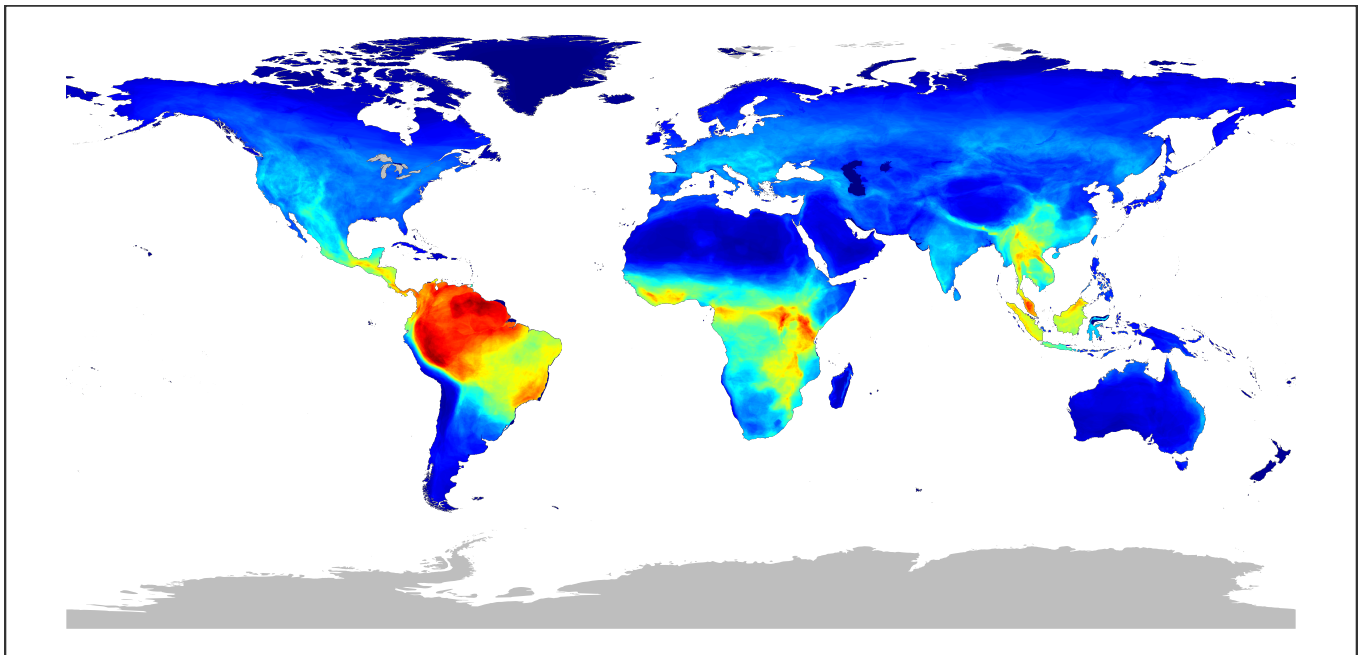


Number of observed susceptible species




0 50 100 150

b) Predicted distribution map for Paramyxoviridae ($p > 0.5$)



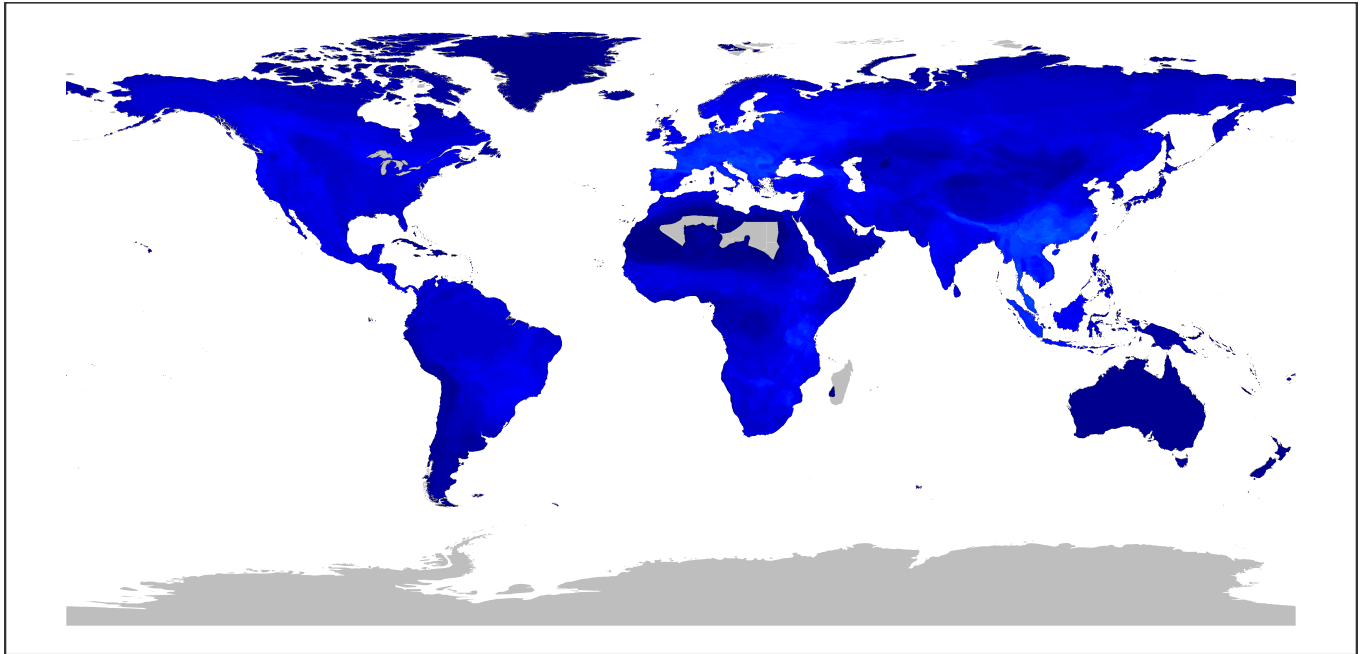
Number of observed and predicted susceptible species




0 50 100 150

Supplementary Figure SR79 | Geographic distribution of associations for Paramyxoviridae. The panels show the observed (a) and predicted (b) richness of mammal species associated with the viral family. A mammal species was considered predicted to be associated with the viral family if the median predicted probability exceeded 0.5.

a) Observed distribution map for Parvoviridae

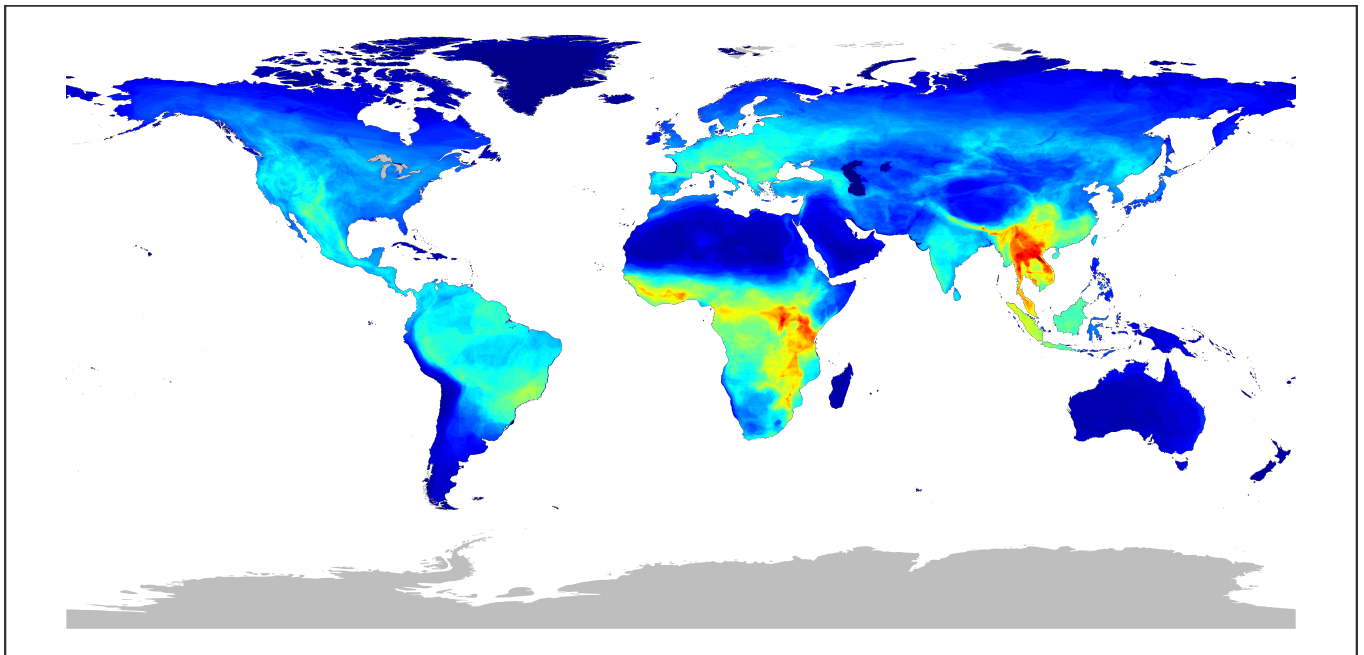


Number of observed susceptible species




0 25 50 75 100 125

b) Predicted distribution map for Parvoviridae ($p > 0.5$)



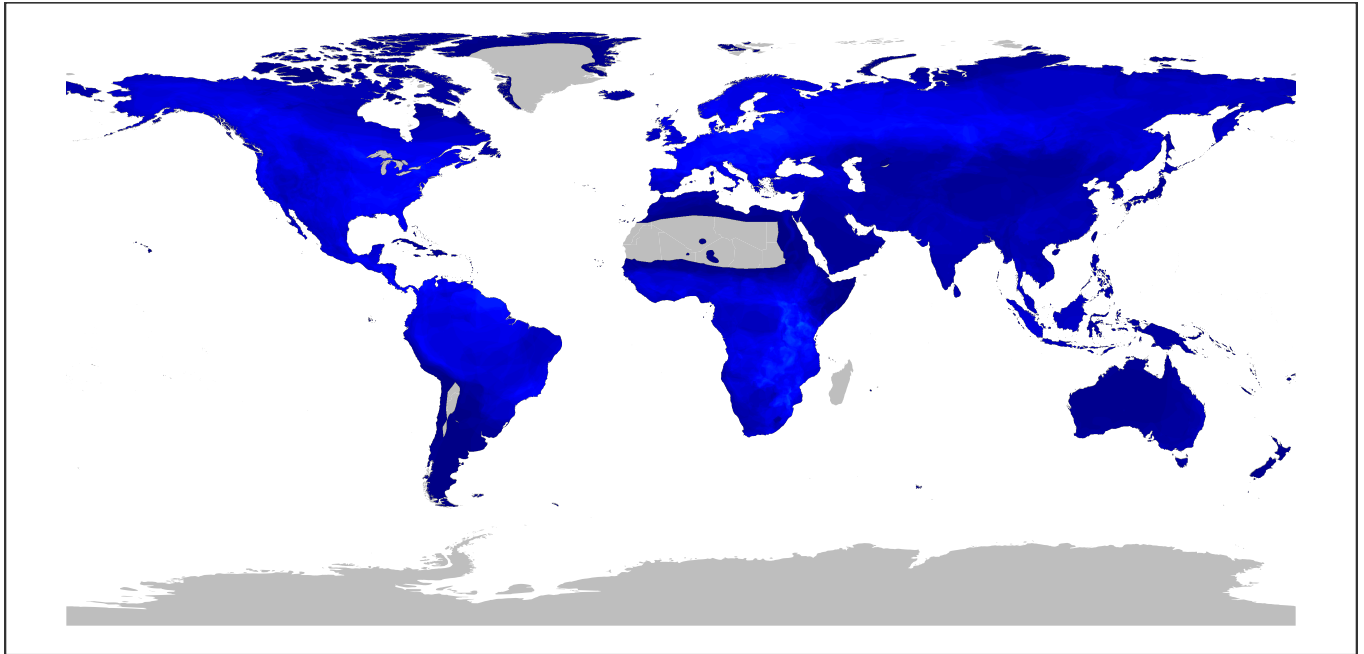
Number of observed and predicted susceptible species




0 25 50 75 100 125

Supplementary Figure SR80 | Geographic distribution of associations for Parvoviridae. The panels show the observed (a) and predicted (b) richness of mammal species associated with the viral family. A mammal species was considered predicted to be associated with the viral family if the median predicted probability exceeded 0.5.

a) Observed distribution map for Peribunyaviridae

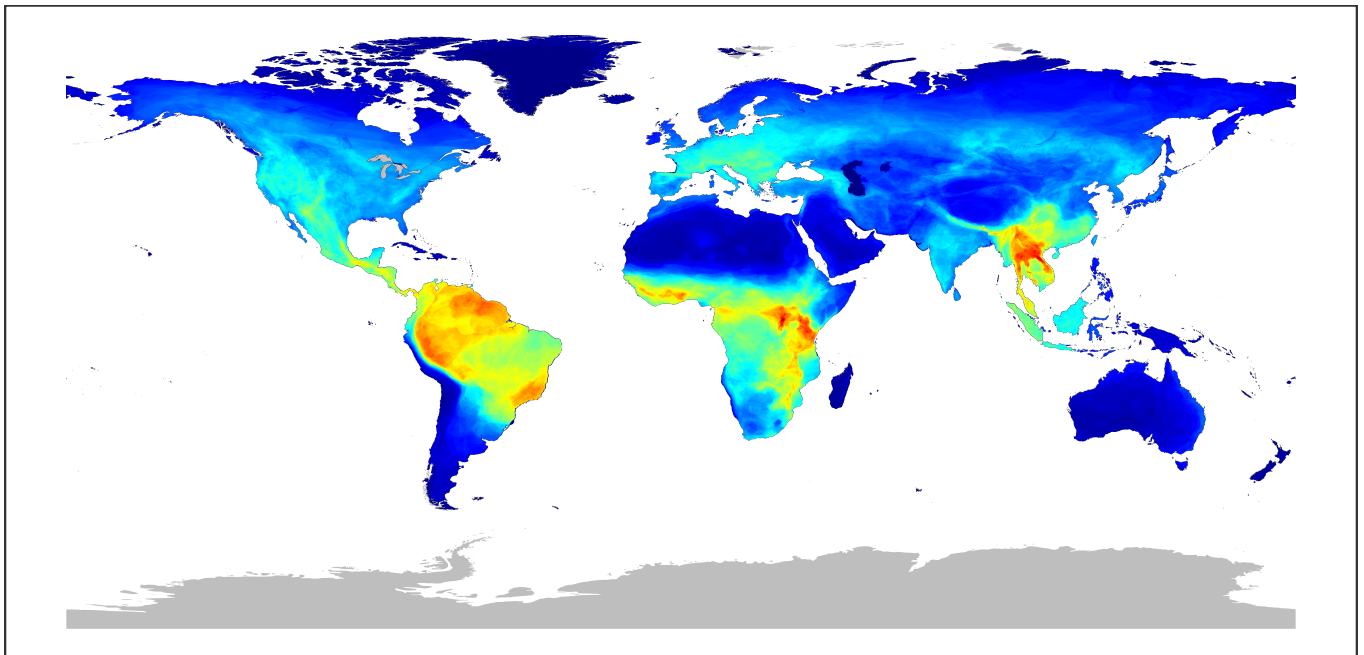


Number of observed susceptible species




0 25 50 75 100 125

b) Predicted distribution map for Peribunyaviridae ($p > 0.5$)



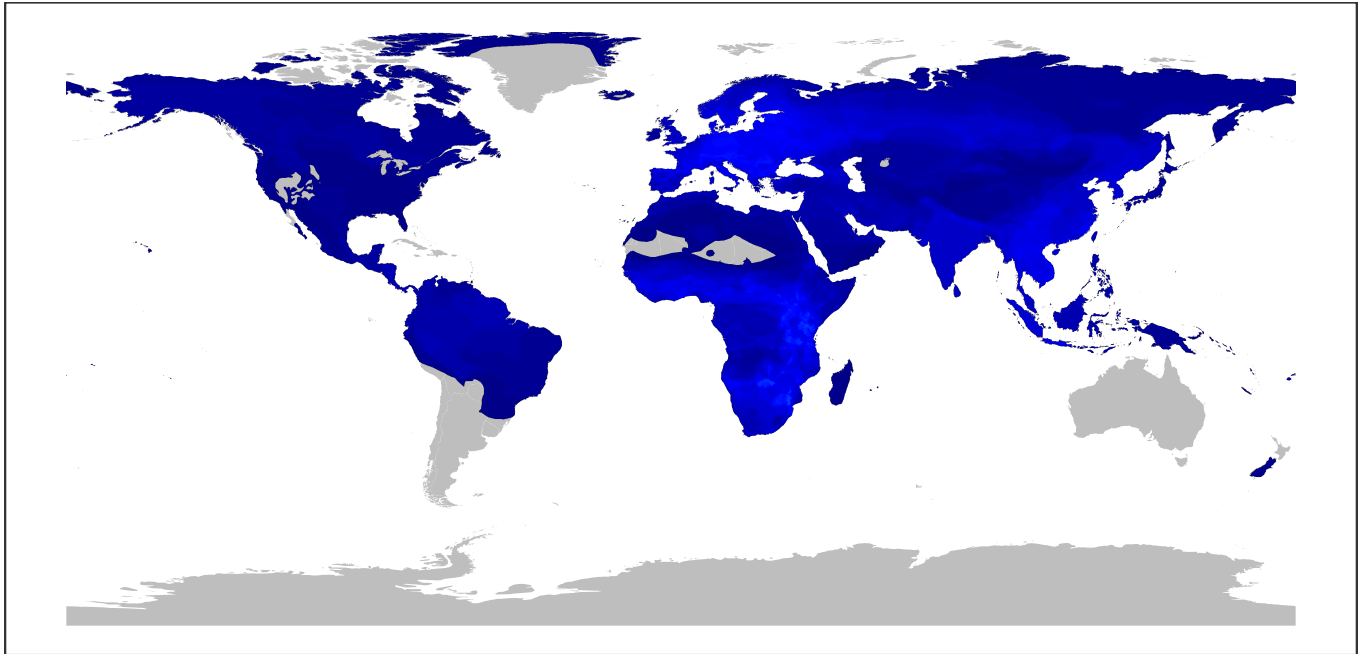
Number of observed and predicted susceptible species




0 25 50 75 100 125

Supplementary Figure SR81 | Geographic distribution of associations for Peribunyaviridae. The panels show the observed (**a**) and predicted (**b**) richness of mammal species associated with the viral family. A mammal species was considered predicted to be associated with the viral family if the median predicted probability exceeded 0.5.

a) Observed distribution map for Phenuiviridae

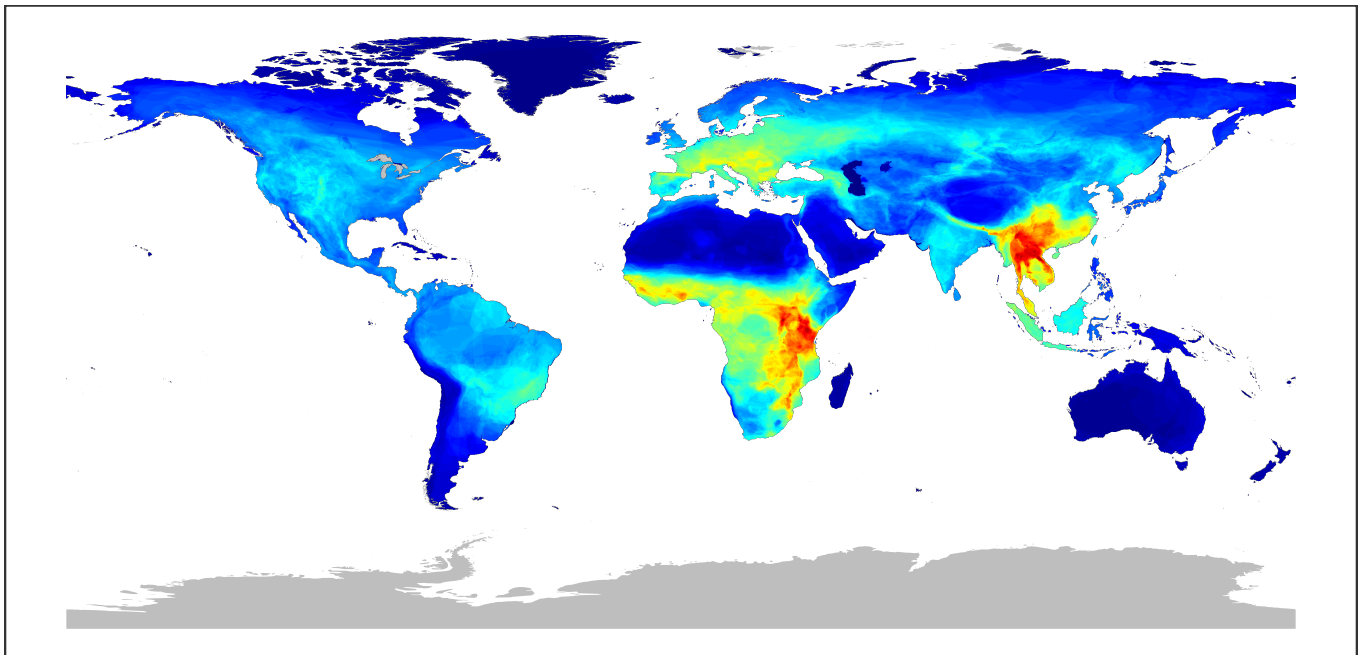


Number of observed susceptible species




0 25 50 75

b) Predicted distribution map for Phenuiviridae ($p > 0.5$)



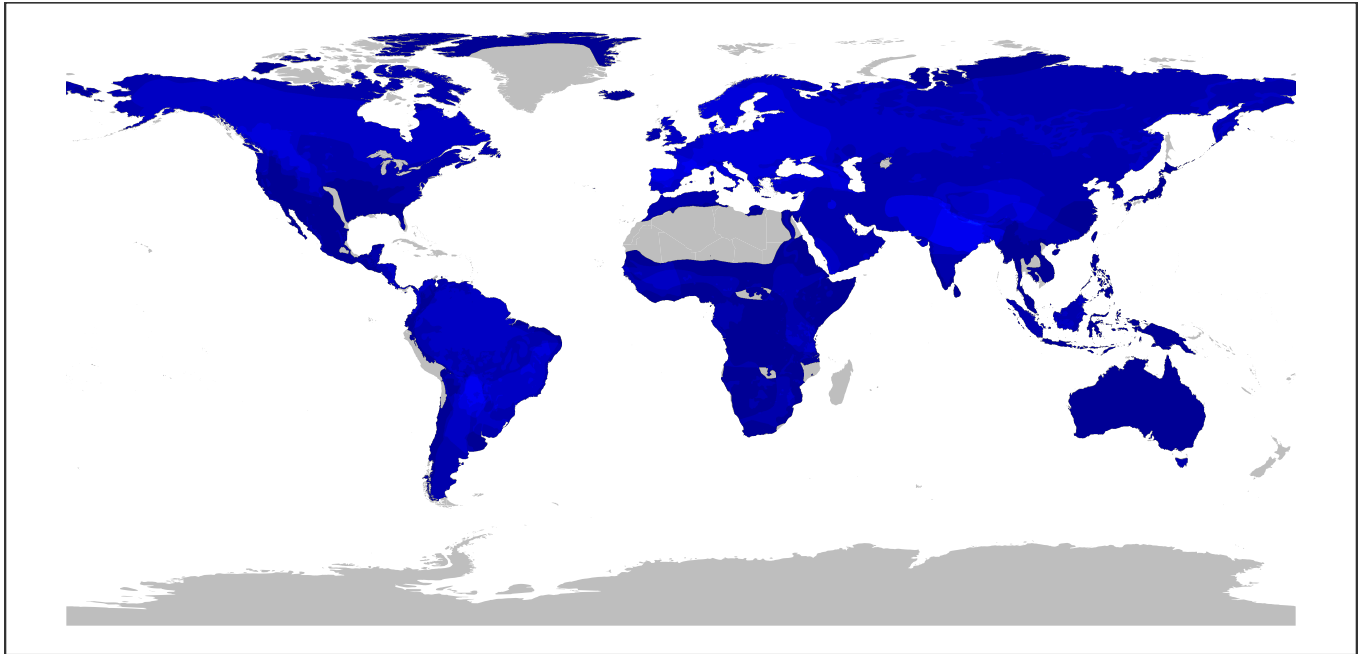
Number of observed and predicted susceptible species




0 25 50 75

Supplementary Figure SR82 | Geographic distribution of associations for Phenuiviridae. The panels show the observed (a) and predicted (b) richness of mammal species associated with the viral family. A mammal species was considered predicted to be associated with the viral family if the median predicted probability exceeded 0.5.

a) Observed distribution map for Picobirnaviridae

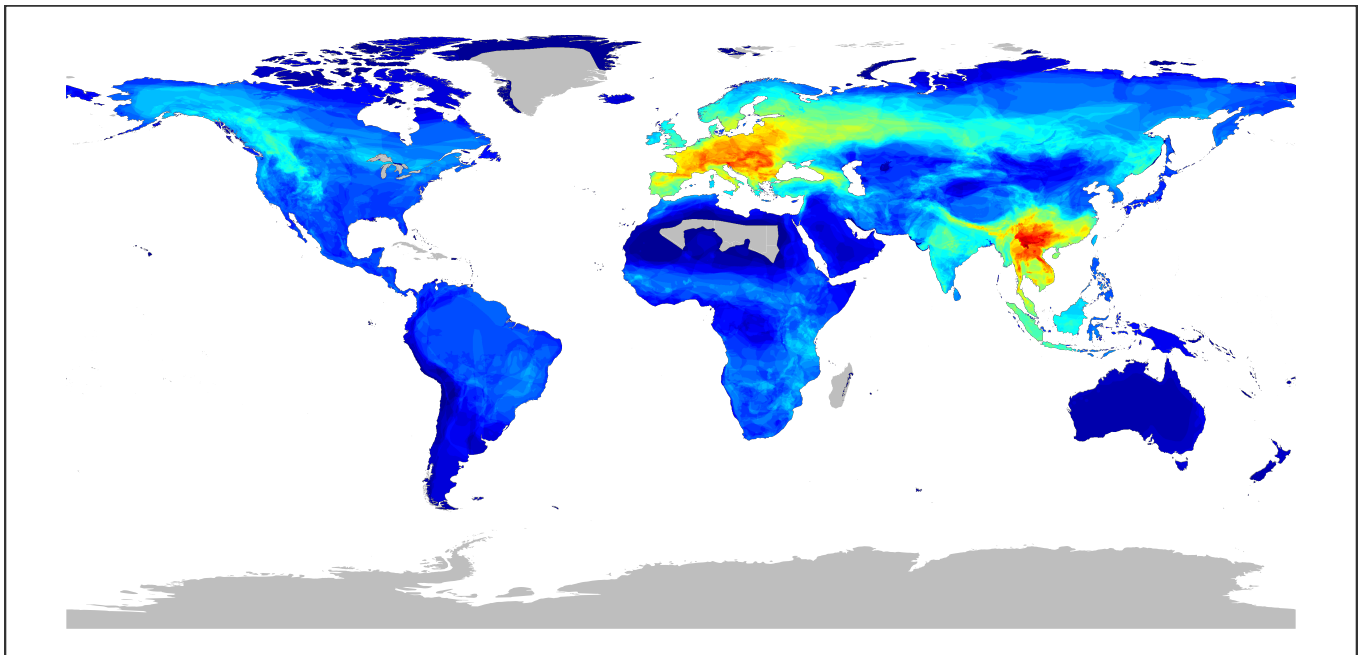


Number of observed susceptible species




0 10 20 30 40

b) Predicted distribution map for Picobirnaviridae ($p > 0.5$)



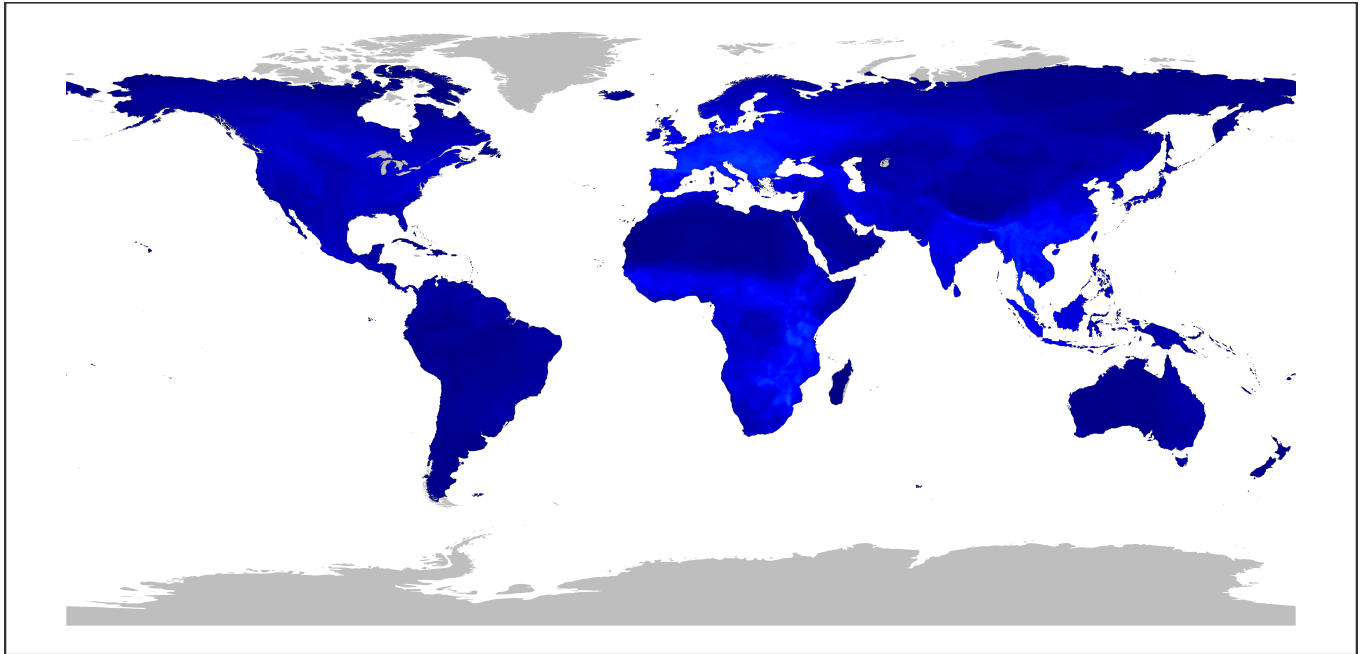
Number of observed and predicted susceptible species




0 10 20 30 40

Supplementary Figure SR83 | Geographic distribution of associations for Picobirnaviridae. The panels show the observed (**a**) and predicted (**b**) richness of mammal species associated with the viral family. A mammal species was considered predicted to be associated with the viral family if the median predicted probability exceeded 0.5.

a) Observed distribution map for Picornaviridae

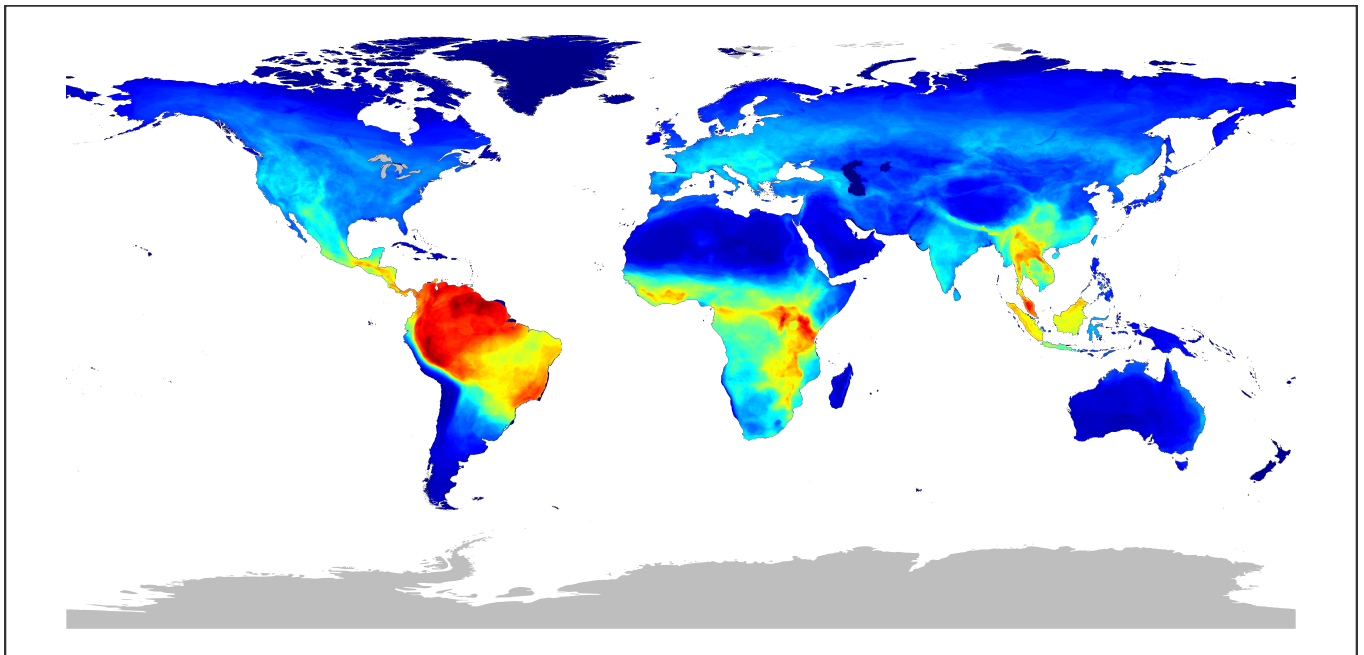


Number of observed susceptible species




0 40 80 120 160

b) Predicted distribution map for Picornaviridae ($p > 0.5$)



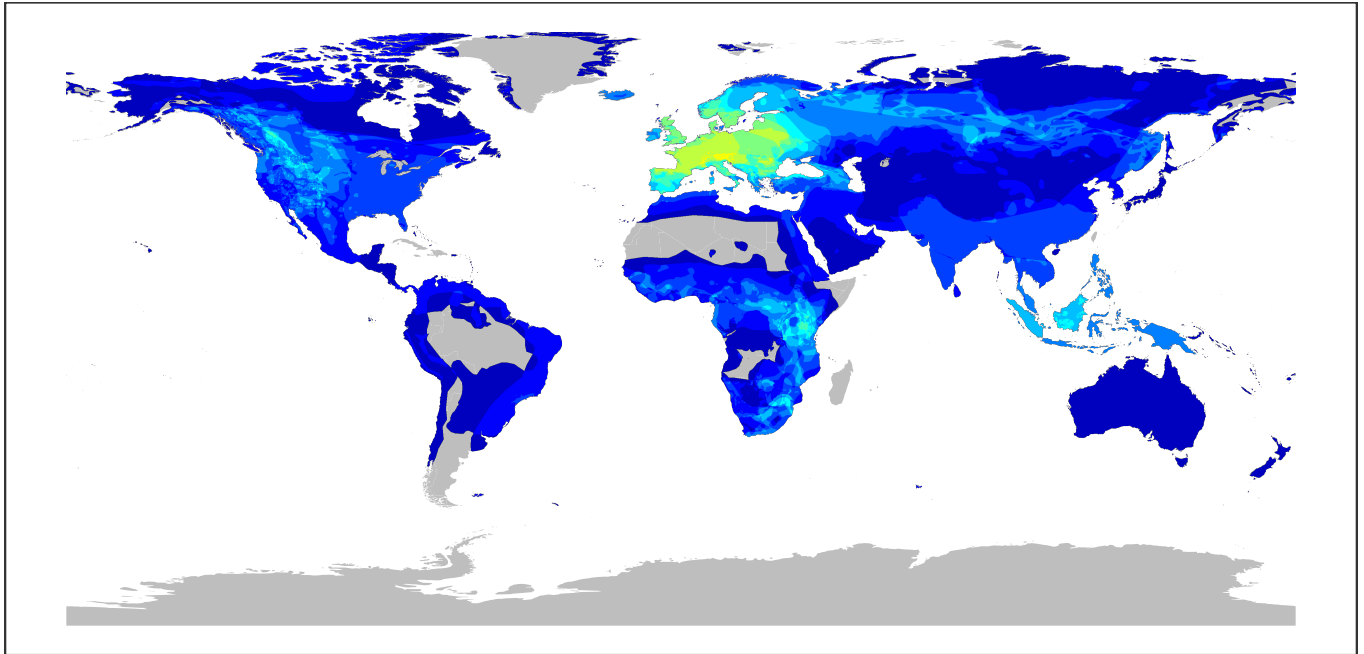
Number of observed and predicted susceptible species




0 40 80 120 160

Supplementary Figure SR84 | Geographic distribution of associations for Picornaviridae. The panels show the observed (**a**) and predicted (**b**) richness of mammal species associated with the viral family. A mammal species was considered predicted to be associated with the viral family if the median predicted probability exceeded 0.5.

a) Observed distribution map for Pneumoviridae

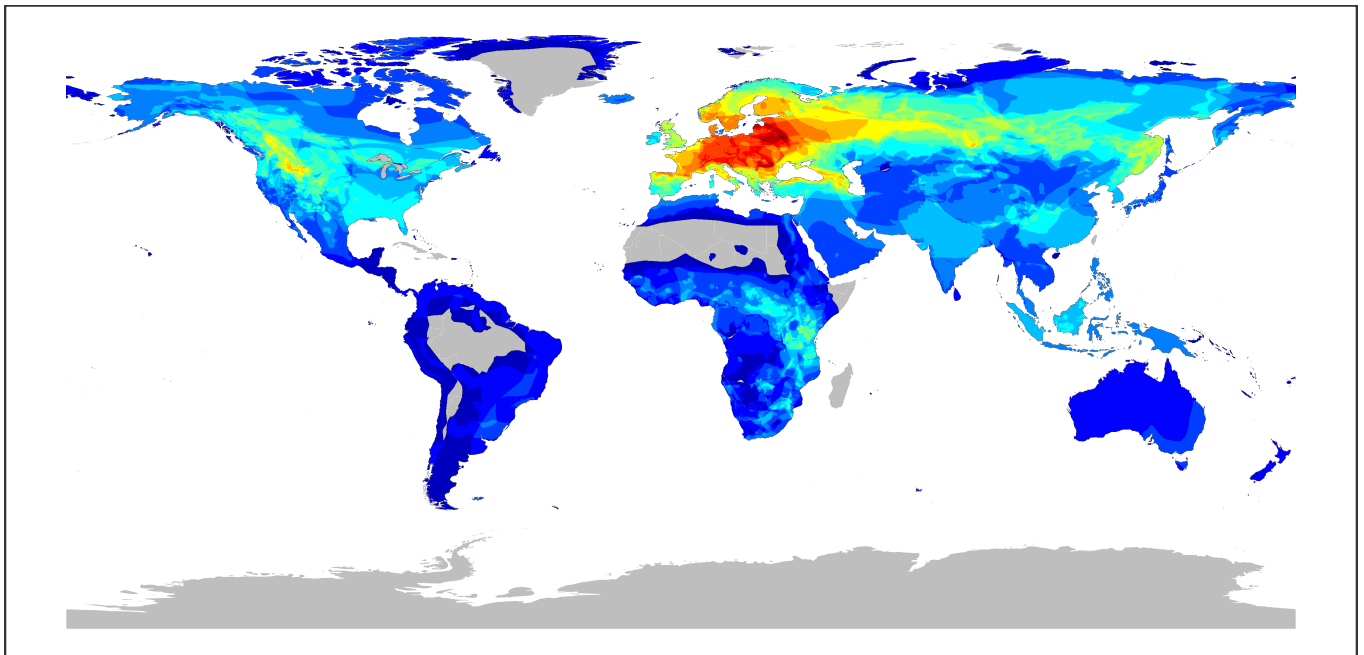


Number of observed susceptible species




0 4 8 12 16

b) Predicted distribution map for Pneumoviridae ($p > 0.5$)



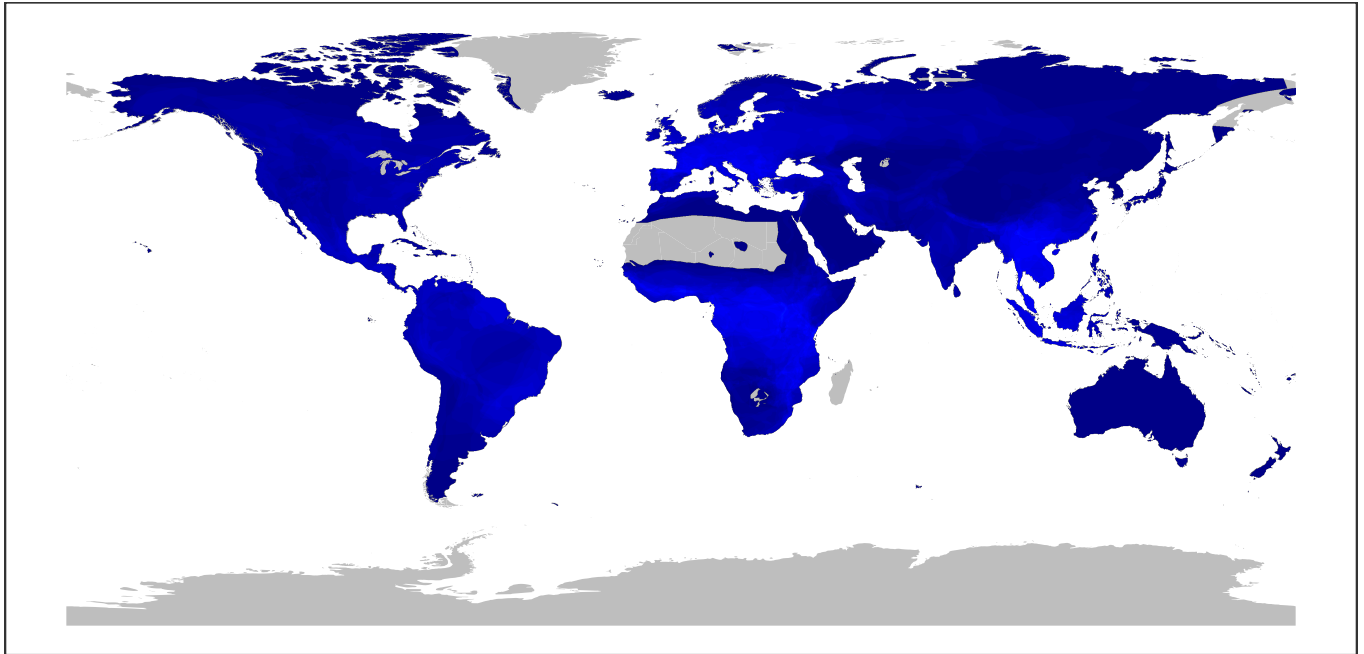
Number of observed and predicted susceptible species




0 4 8 12 16

Supplementary Figure SR85 | Geographic distribution of associations for Pneumoviridae. The panels show the observed (a) and predicted (b) richness of mammal species associated with the viral family. A mammal species was considered predicted to be associated with the viral family if the median predicted probability exceeded 0.5.

a) Observed distribution map for Polyomaviridae

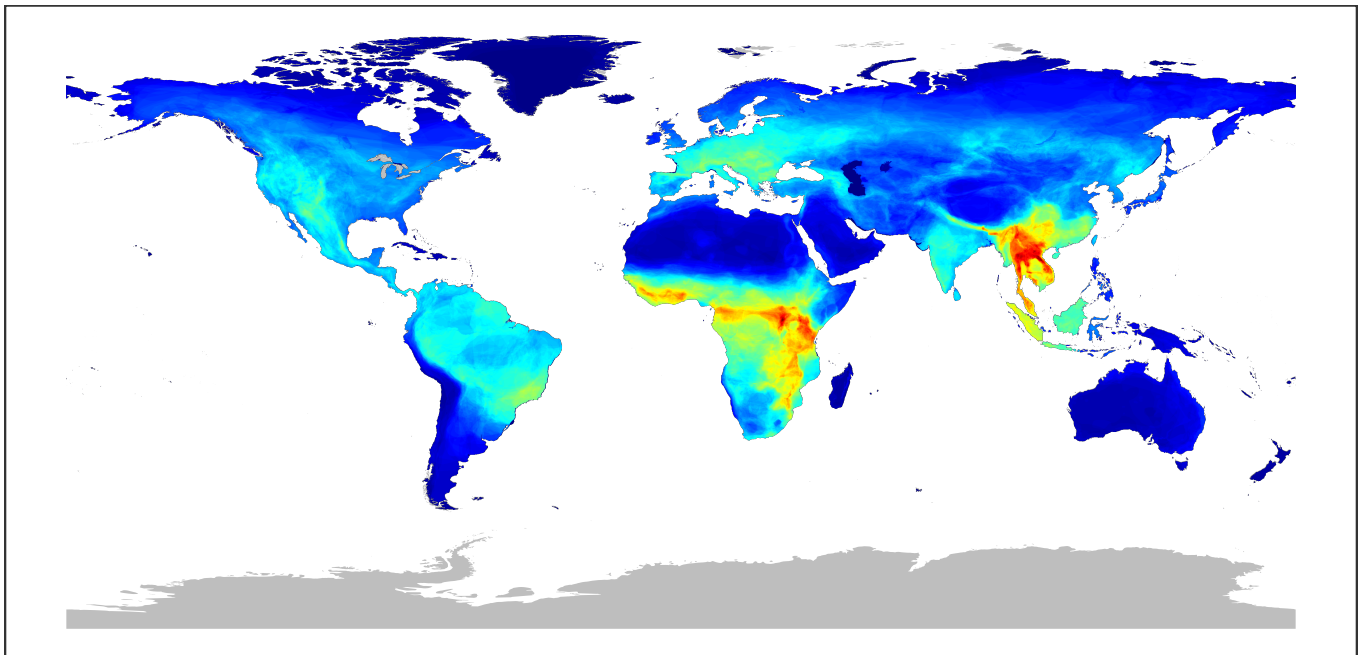


Number of observed susceptible species




0 30 60 90 120

b) Predicted distribution map for Polyomaviridae ($p > 0.5$)



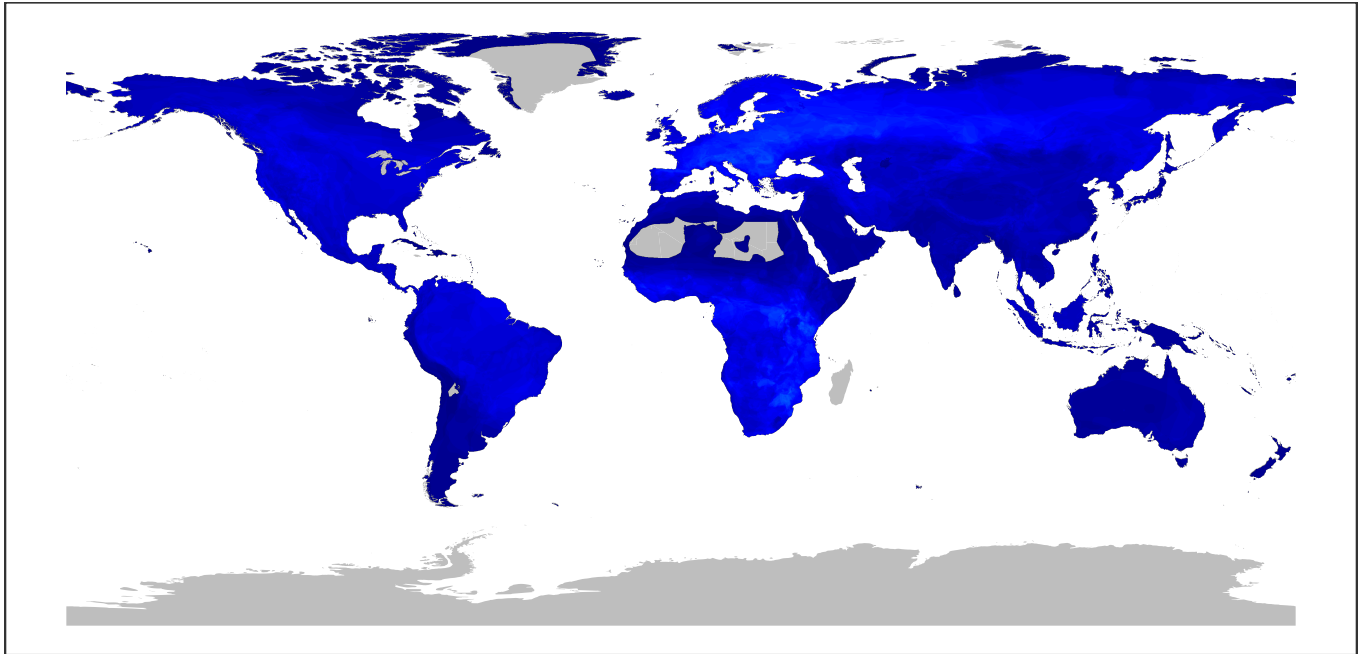
Number of observed and predicted susceptible species




0 30 60 90 120

Supplementary Figure SR86 | Geographic distribution of associations for Polyomaviridae. The panels show the observed (**a**) and predicted (**b**) richness of mammal species associated with the viral family. A mammal species was considered predicted to be associated with the viral family if the median predicted probability exceeded 0.5.

a) Observed distribution map for Poxviridae

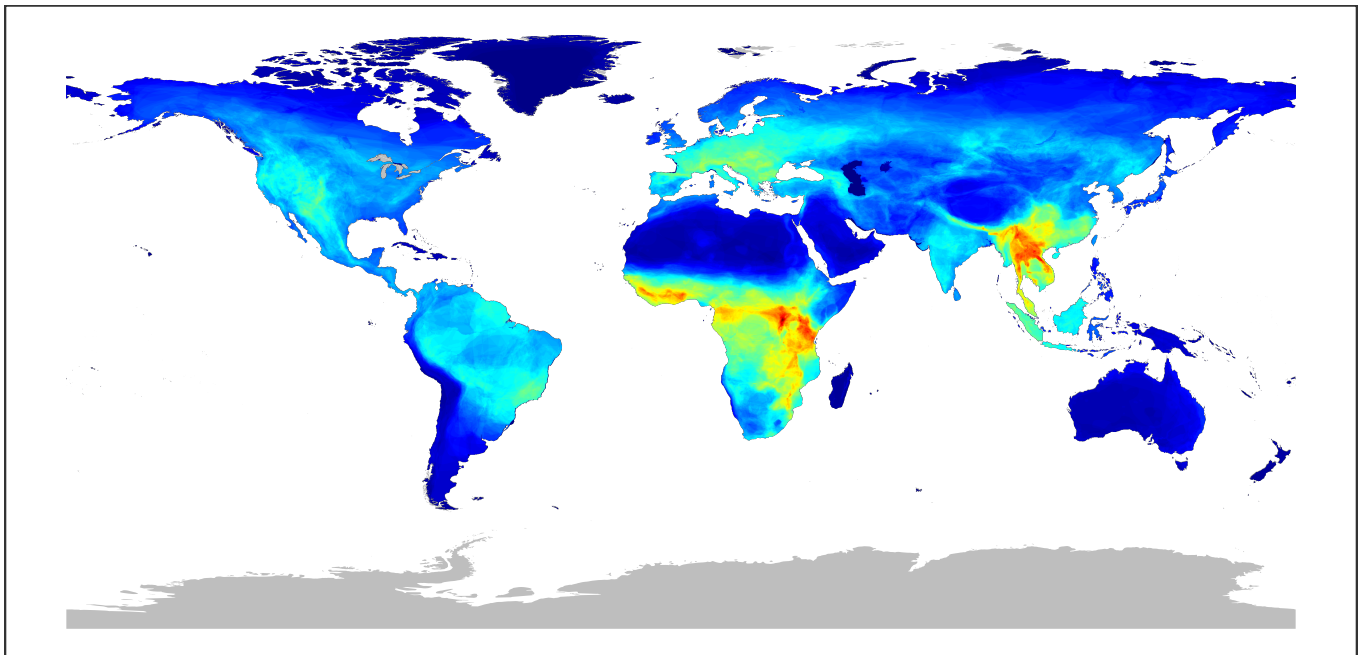


Number of observed susceptible species




0 30 60 90

b) Predicted distribution map for Poxviridae ($p > 0.5$)



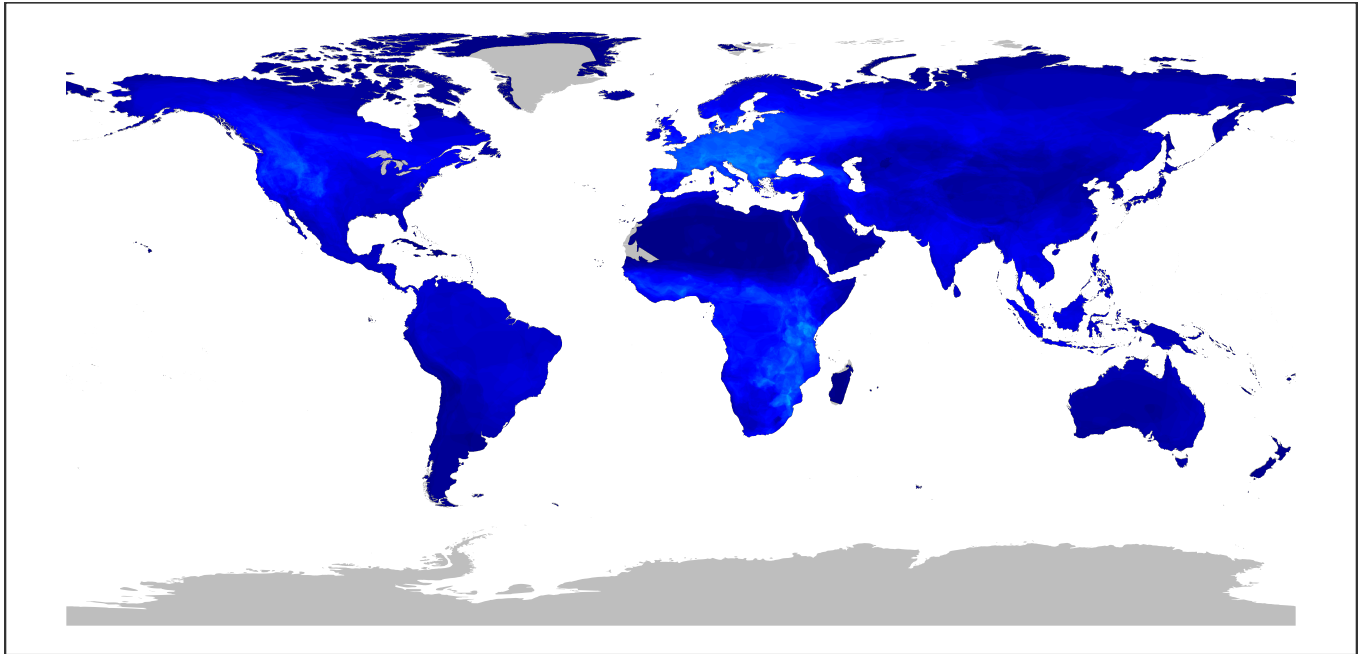
Number of observed and predicted susceptible species




0 30 60 90

Supplementary Figure SR87 | Geographic distribution of associations for Poxviridae. The panels show the observed (**a**) and predicted (**b**) richness of mammal species associated with the viral family. A mammal species was considered predicted to be associated with the viral family if the median predicted probability exceeded 0.5.

a) Observed distribution map for Reoviridae

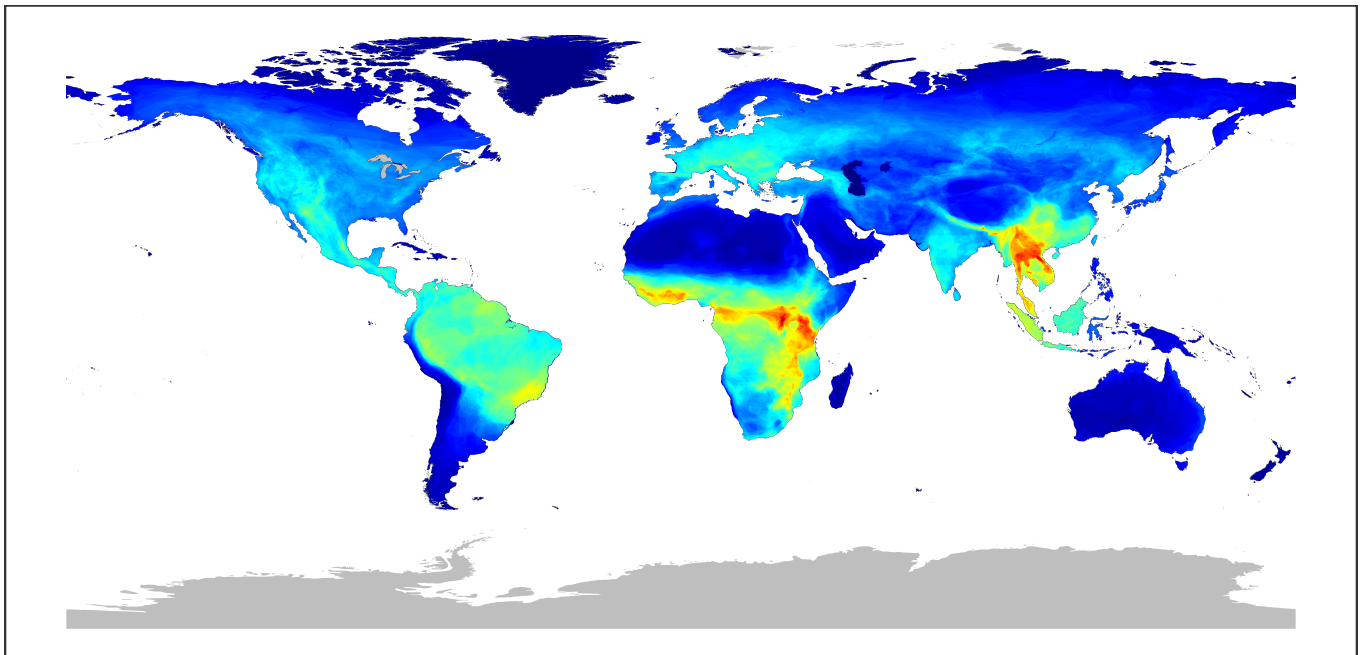


Number of observed susceptible species




0 50 100

b) Predicted distribution map for Reoviridae ($p > 0.5$)



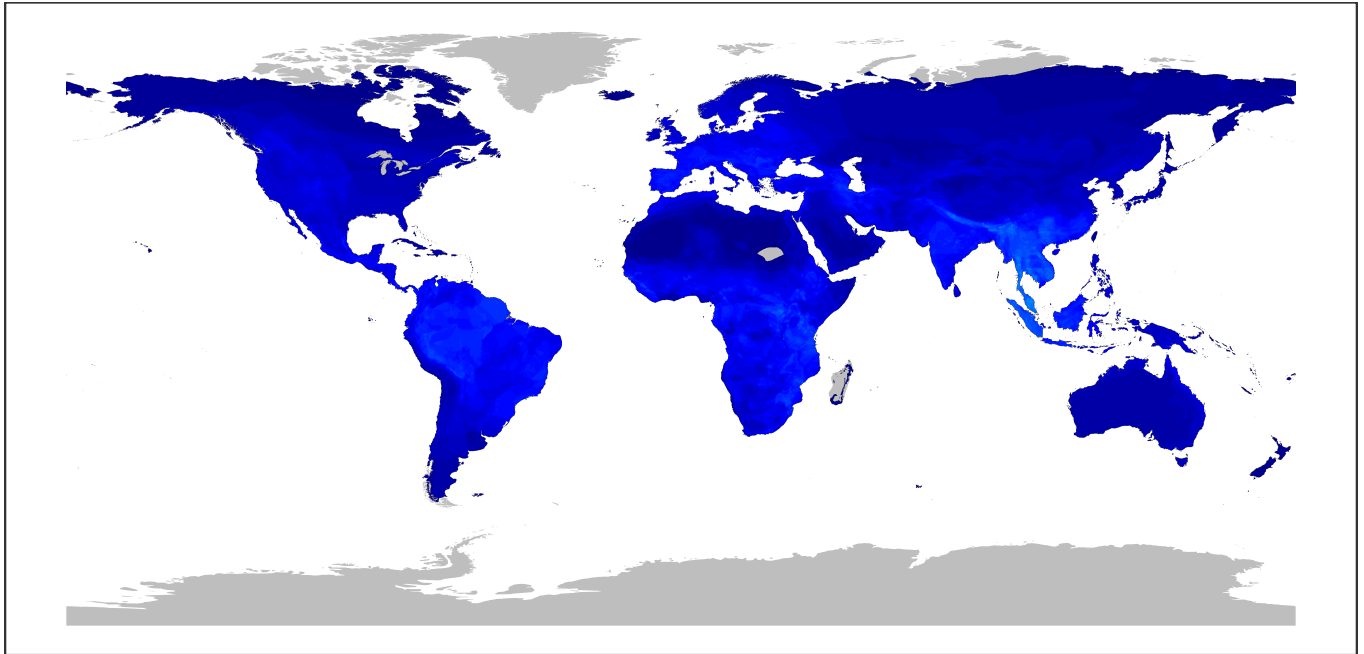
Number of observed and predicted susceptible species




0 50 100

Supplementary Figure SR88 | Geographic distribution of associations for Reoviridae. The panels show the observed (a) and predicted (b) richness of mammal species associated with the viral family. A mammal species was considered predicted to be associated with the viral family if the median predicted probability exceeded 0.5.

a) Observed distribution map for Retroviridae

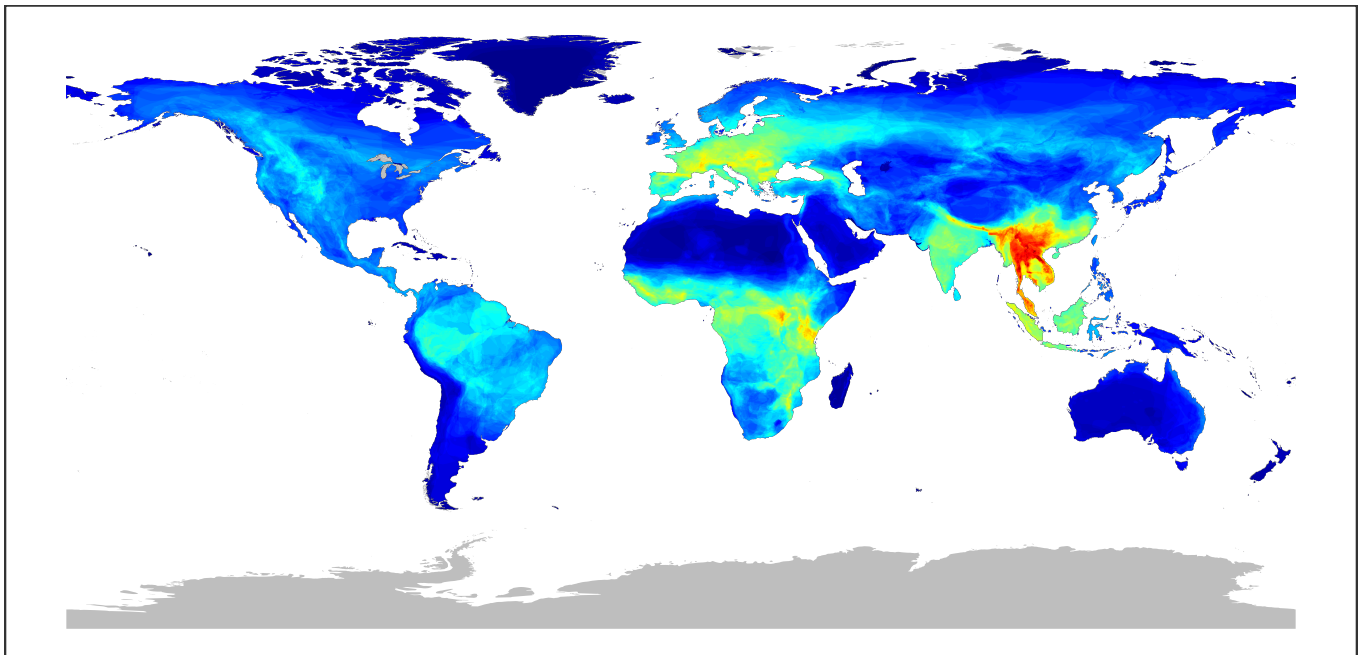


Number of observed susceptible species




0 20 40 60

b) Predicted distribution map for Retroviridae ($p > 0.5$)



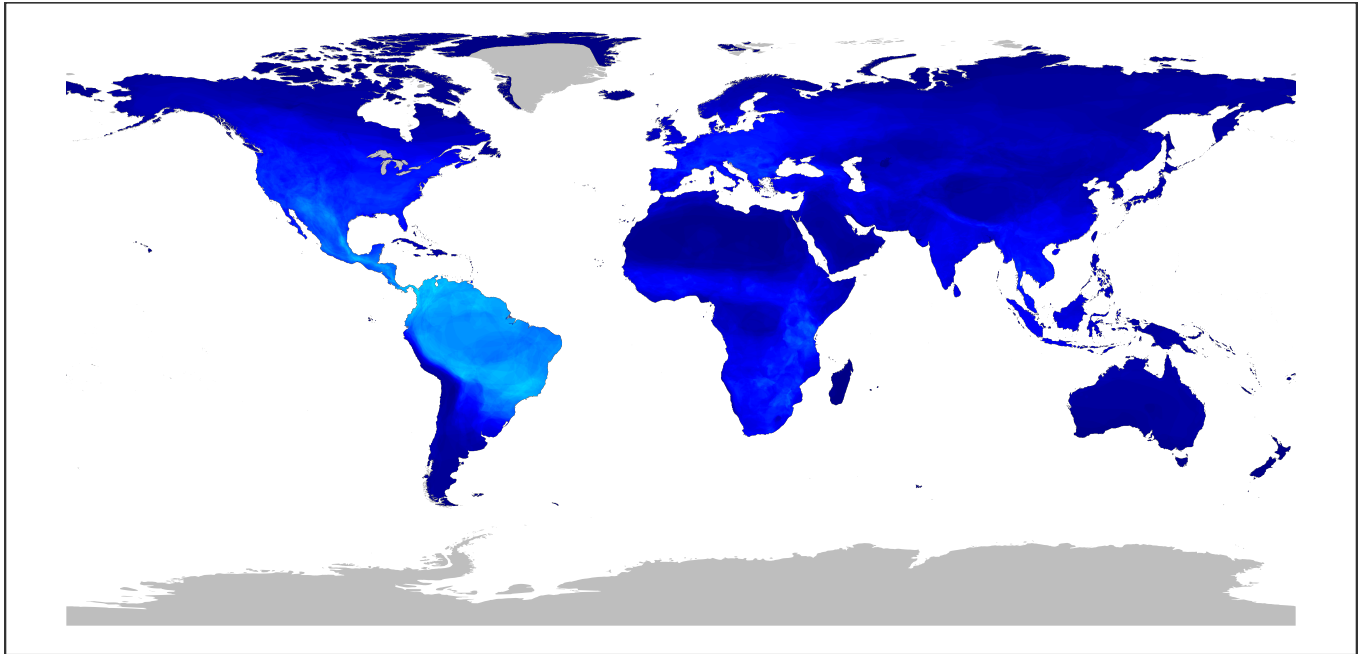
Number of observed and predicted susceptible species




0 20 40 60

Supplementary Figure SR89 | Geographic distribution of associations for Retroviridae. The panels show the observed (**a**) and predicted (**b**) richness of mammal species associated with the viral family. A mammal species was considered predicted to be associated with the viral family if the median predicted probability exceeded 0.5.

a) Observed distribution map for Rhabdoviridae

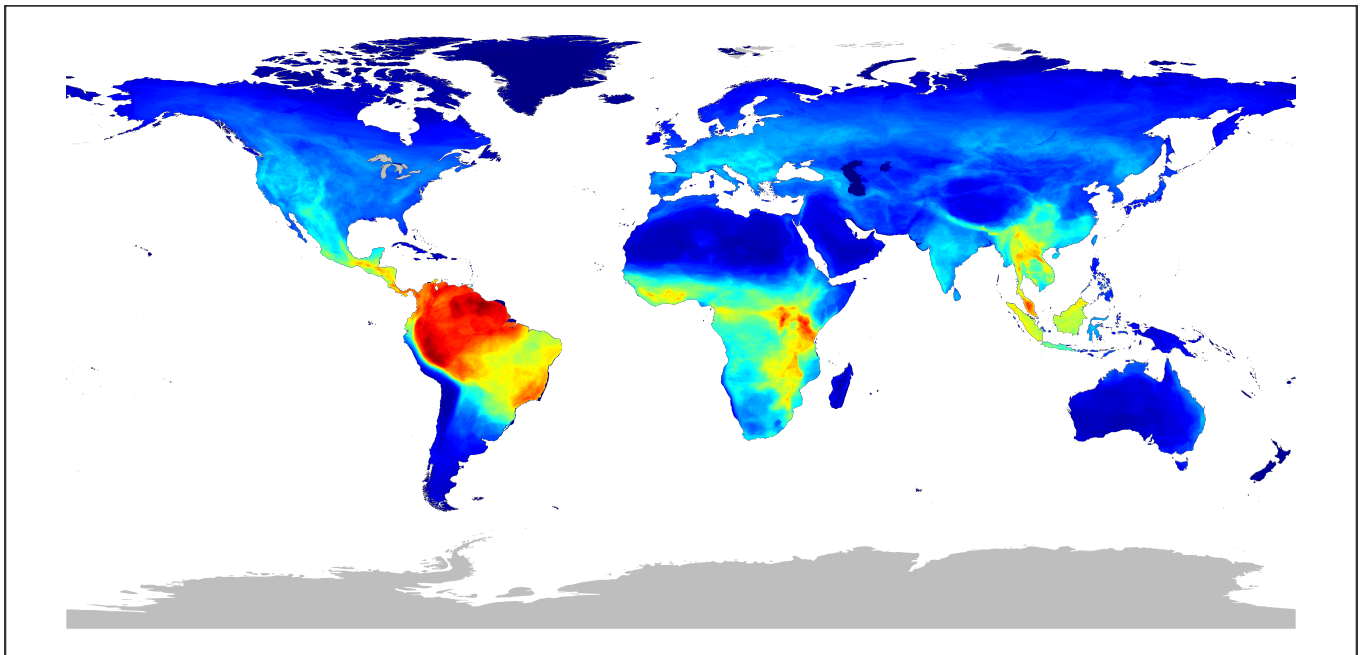


Number of observed susceptible species




0 50 100 150

b) Predicted distribution map for Rhabdoviridae ($p > 0.5$)



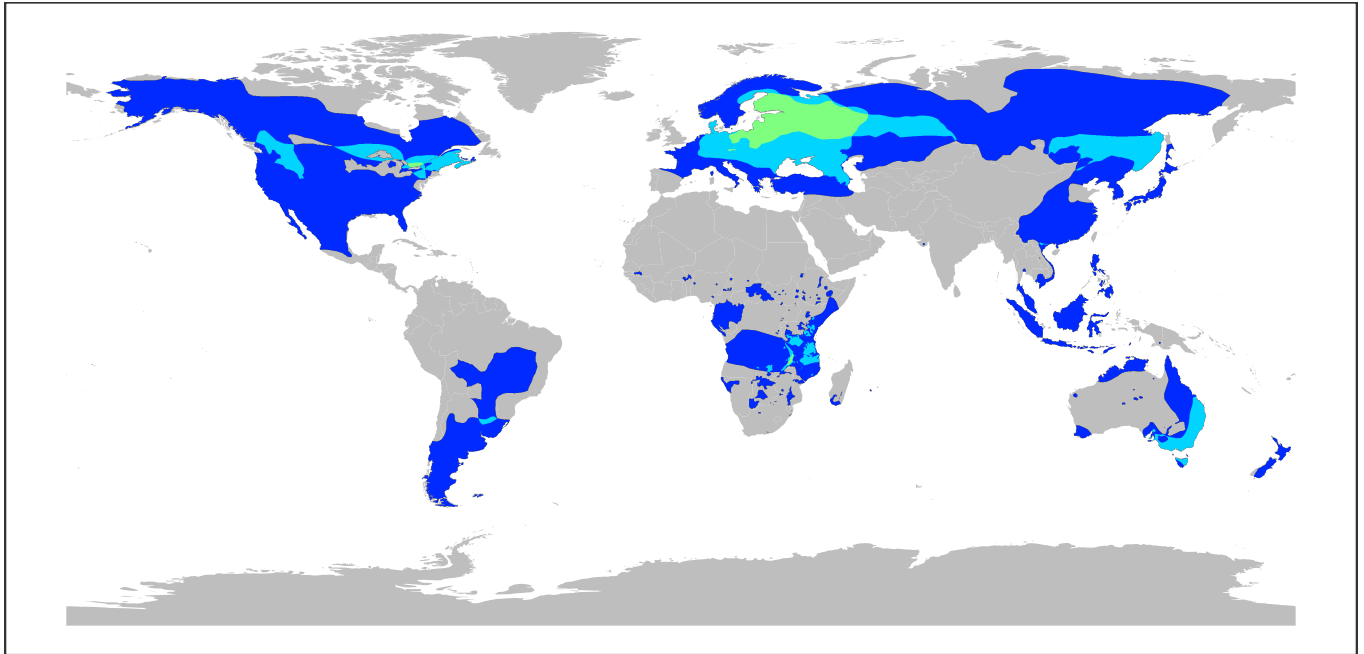
Number of observed and predicted susceptible species




0 50 100 150

Supplementary Figure SR90 | Geographic distribution of associations for Rhabdoviridae. The panels show the observed (**a**) and predicted (**b**) richness of mammal species associated with the viral family. A mammal species was considered predicted to be associated with the viral family if the median predicted probability exceeded 0.5.

a) Observed distribution map for Smacoviridae

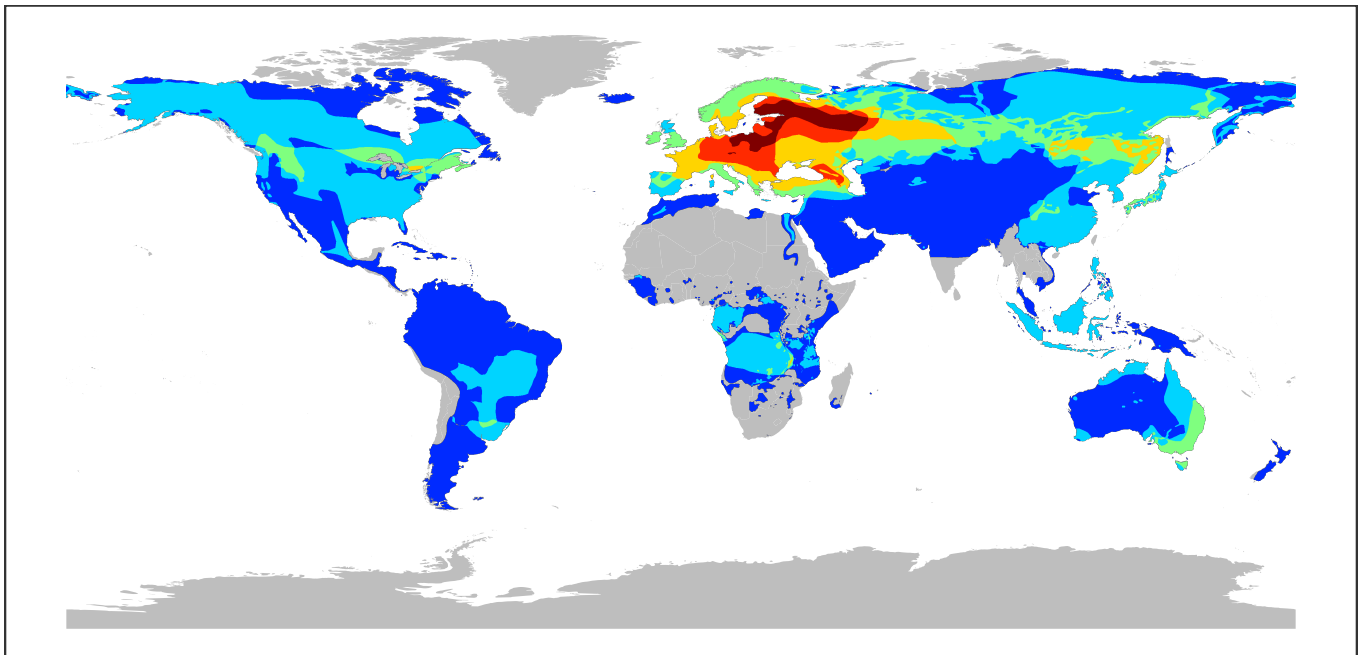


Number of observed susceptible species




0 1 2 3 4 5 6

b) Predicted distribution map for Smacoviridae ($p > 0.5$)



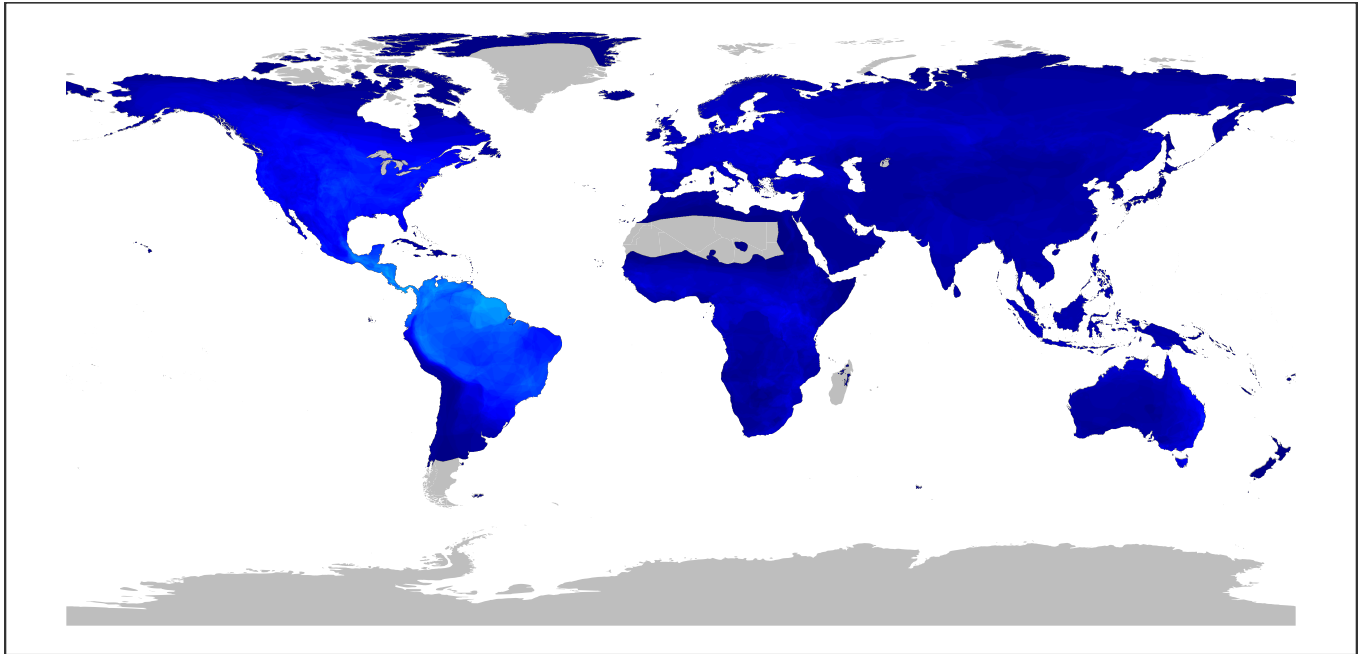
Number of observed and predicted susceptible species




0 1 2 3 4 5 6

Supplementary Figure SR91 | Geographic distribution of associations for Smacoviridae. The panels show the observed (**a**) and predicted (**b**) richness of mammal species associated with the viral family. A mammal species was considered predicted to be associated with the viral family if the median predicted probability exceeded 0.5.

a) Observed distribution map for Togaviridae

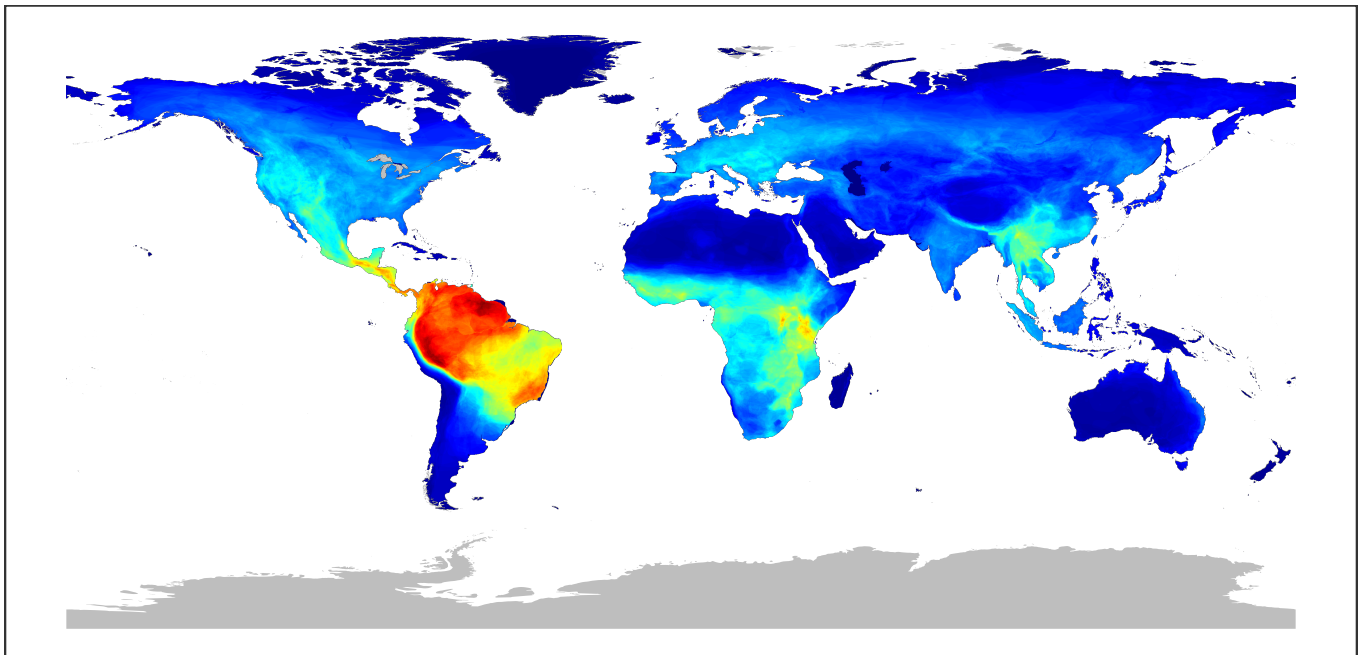


Number of observed susceptible species




0 50 100

b) Predicted distribution map for Togaviridae ($p > 0.5$)



Number of observed and predicted susceptible species

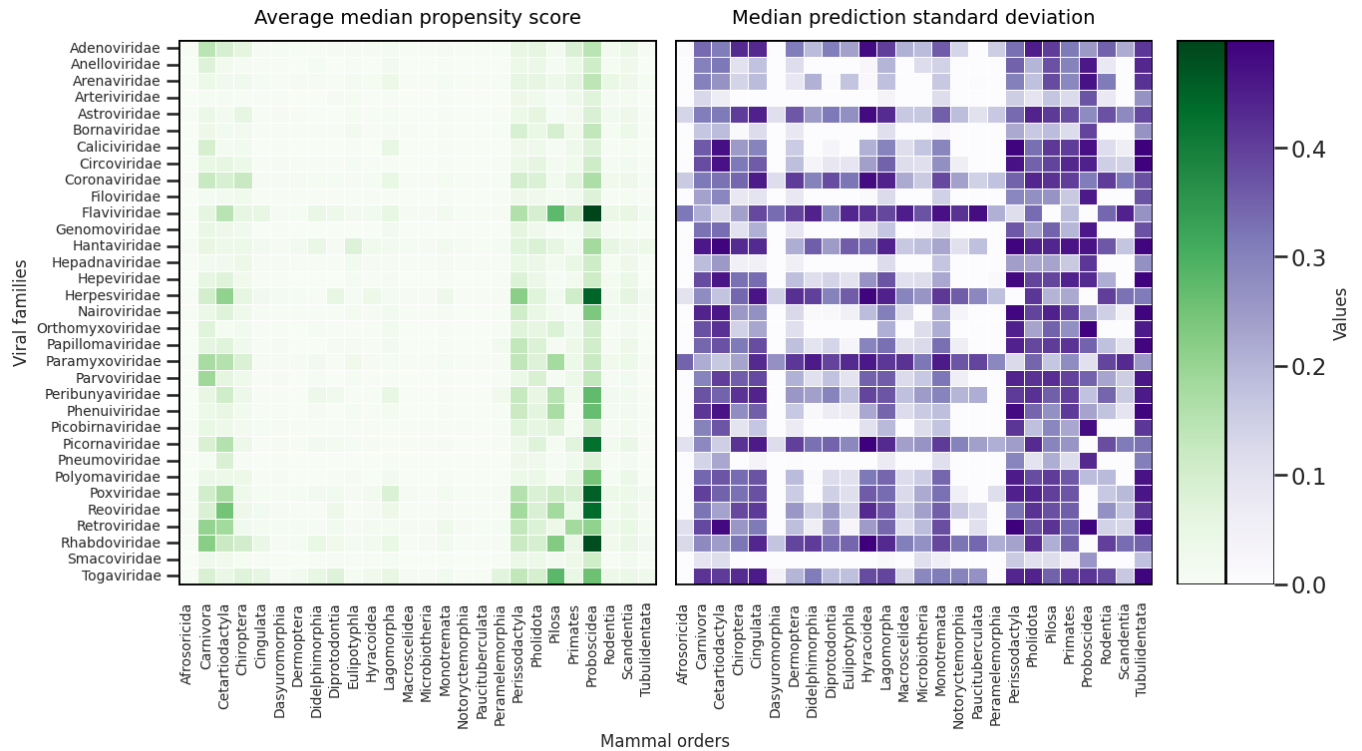


0 50 100

Supplementary Figure SR92 | Geographic distribution of associations for Togaviridae. The panels show the observed (**a**) and predicted (**b**) richness of mammal species associated with the viral family. A mammal species was considered predicted to be associated with the viral family if the median predicted probability exceeded 0.5.

3 Propensity scores and uncertainties

Supplementary Figure SR93 shows mean propensity scores and median uncertainties for all different couples of mammalian orders and viral families. This visualization helps identify taxonomic groups where model predictions are more uncertain and reveals where research efforts have been more concentrated.



Supplementary Figure SR93 | Predicted probabilities and uncertainties across mammalian orders and viral families. Heatmap showing the mean propensity score (left) and the median uncertainty (right) for each mammalian order–viral family pair. Median uncertainty is calculated as the median standard deviation of predictions across all potential associations within each order–family combination.

Also, we report in Supplementary Table SR1 and Supplementary Table SR2 the average propensity score and prediction uncertainty across mammalian orders and viral families, respectively.

Supplementary Table SR1 | Predicted propensity scores and prediction uncertainty across mammalian orders. a) For each mammalian order, the table reports the mean of the average median propensity score across viral families. b) For each mammalian order, the table reports the mean of the median standard deviation of predictions across viral families.

a) Propensity Score				b) Prediction Uncertainty (Std. Dev.)			
Order	Score	Order	Score	Order	Std Dev	Order	Std Dev
Proboscidea	0.313	Perissodactyla	0.136	Tubulidentata	0.403	Perissodactyla	0.364
Cetartiodactyla	0.122	Carnivora	0.122	Cetartiodactyla	0.339	Pholidota	0.330
Pilosa	0.106	Pholidota	0.097	Pilosa	0.329	Primates	0.328
Primates	0.072	Chiroptera	0.060	Carnivora	0.326	Proboscidea	0.304
Scandentia	0.050	Lagomorpha	0.036	Lagomorpha	0.302	Cingulata	0.297
Rodentia	0.030	Diprotodontia	0.020	Chiroptera	0.277	Hyracoidea	0.268
Didelphimorphia	0.020	Eulipotyphla	0.019	Monotremata	0.266	Dermoptera	0.219
Cingulata	0.018	Hyracoidea	0.015	Rodentia	0.211	Diprotodontia	0.156
Monotremata	0.015	Tubulidentata	0.014	Didelphimorphia	0.142	Microbiotheria	0.142
Peramelemorphia	0.013	Dasyuromorphia	0.011	Eulipotyphla	0.138	Scandentia	0.128
Dermoptera	0.008	Microbiotheria	0.006	Macroscelidea	0.121	Notoryctemorphia	0.099
Macroscelidea	0.005	Afrosoricida	0.005	Paucituberculata	0.087	Peramelemorphia	0.048
Notoryctemorphia	0.004	Paucituberculata	0.004	Afrosoricida	0.043	Dasyuromorphia	0.038

Supplementary Table SR2 | Predicted propensity scores and uncertainty across viral families. **a)** For each viral family, the table reports the mean of the average median propensity score across mammalian orders. **b)** For each viral family, the table reports the mean of the median standard deviation of predictions across mammalian orders.

a) Propensity Score

Viral Family	Score	Viral Family	Score
Flaviviridae	0.110	Rhabdoviridae	0.109
Herpesviridae	0.095	Togaviridae	0.095
Poxviridae	0.093	Reoviridae	0.093
Retroviridae	0.081	Paramyxoviridae	0.072
Picornaviridae	0.065	Peribunyaviridae	0.063
Coronaviridae	0.061	Hantaviridae	0.060
Phenuiviridae	0.051	Adenoviridae	0.051
Nairoviridae	0.043	Parvoviridae	0.042
Arenaviridae	0.041	Papillomaviridae	0.040
Polyomaviridae	0.038	Bornaviridae	0.036
Orthomyxoviridae	0.033	Hepeviridae	0.032
Picobirnaviridae	0.032	Hepadnaviridae	0.031
Circoviridae	0.030	Astroviridae	0.028
Anelloviridae	0.027	Caliciviridae	0.025
Smacoviridae	0.024	Genomoviridae	0.022
Pneumoviridae	0.020	Filoviridae	0.019
Arteriviridae	0.018		

b) Prediction Uncertainty (Std. Dev.)

Viral Family	Std Dev	Viral Family	Std Dev
Paramyxoviridae	0.337	Flaviviridae	0.323
Coronaviridae	0.323	Rhabdoviridae	0.312
Astroviridae	0.307	Hantaviridae	0.305
Togaviridae	0.300	Picornaviridae	0.299
Herpesviridae	0.295	Adenoviridae	0.281
Peribunyaviridae	0.281	Retroviridae	0.246
Reoviridae	0.244	Parvoviridae	0.241
Hepeviridae	0.233	Circoviridae	0.232
Papillomaviridae	0.229	Poxviridae	0.226
Polyomaviridae	0.221	Caliciviridae	0.211
Phenuiviridae	0.210	Nairoviridae	0.205
Arenaviridae	0.175	Orthomyxoviridae	0.167
Anelloviridae	0.153	Genomoviridae	0.145
Picobirnaviridae	0.143	Filoviridae	0.122
Hepadnaviridae	0.114	Pneumoviridae	0.106
Bornaviridae	0.091	Arteriviridae	0.090
Smacoviridae	0.074		

4 Predictions by mammal degrees

Supplementary Figure SR94 shows how model predictions vary with the number of known viral family associations per host species. The results indicate two main trends. First, species with few known associations are predicted to have proportionally more undiscovered links—an expected, though non-trivial, outcome. Second, species with many known associations tend to receive higher predicted probabilities for their unlabeled links, while species with few or no known associations (e.g., fewer than five) tend to receive lower probabilities.

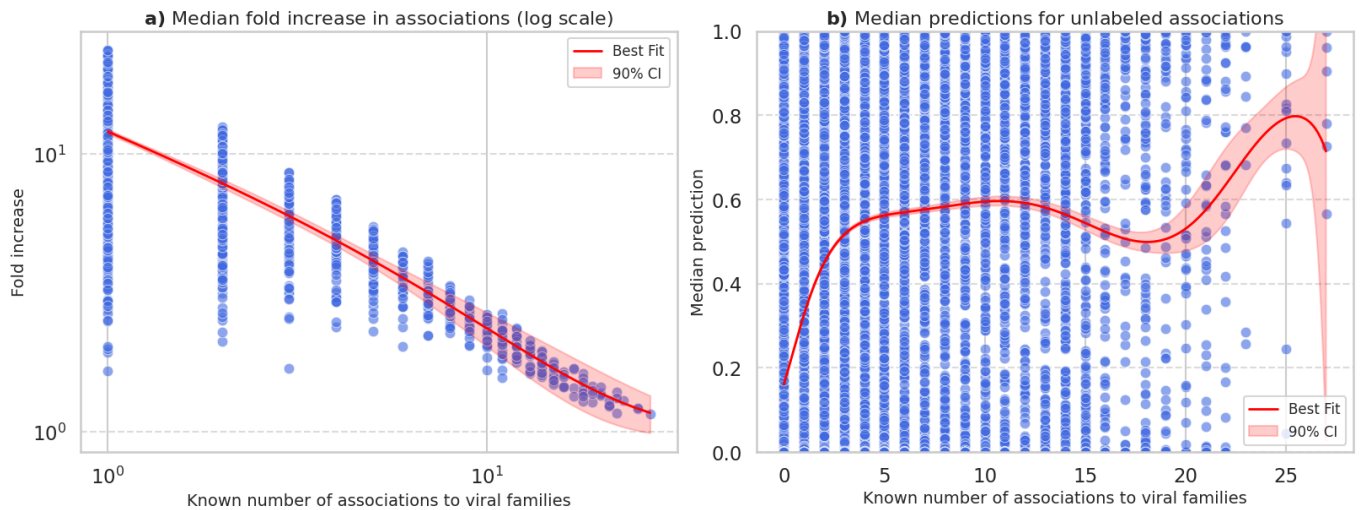
This pattern may arise for several reasons. One possibility is that the model tends to be more confident in making positive predictions for well-sampled species because their better-characterized viral profiles offer clearer patterns from which the model can extrapolate. Another explanation is the presence of residual bias favoring data-rich hosts, whereby species with more known associations are predicted with higher probabilities simply because they are more frequently observed—not necessarily due to underlying ecological or biological factors. Alternatively, the trend may reflect a genuine ecological signal: species with many observed associations might indeed harbor more viruses, potentially due to broader geographic ranges, intrinsic biological traits, or increased contact with other species—all factors that could increase their exposure to a wider diversity of viral families. In contrast, understudied species tend to be more isolated—geographically, phylogenetically, or both—and may therefore have fewer opportunities for viral transmission involving the specific viral families considered in this study.

Nonetheless, overall, the predicted probabilities for unlabeled associations do not appear to depend strongly on the number of known associations.

5 Table of predictions

inally, in Extended Data Table 2, we report the predictions for all pairs of mammal species and viral families $(m, v) \in M \times V$. Each row corresponds to a potential mammal–virus pair (m, v) , with statistical summaries derived from ensemble model predictions:

1. **viral_family**: The name of the viral family.
2. **iucn2020_binomial**: The scientific (binomial) name of the mammal species according to the IUCN 2020 taxonomy.
3. **order**: The taxonomic order of the mammal species.
4. **family**: The taxonomic family of the mammal species.
5. **label**: A binary or categorical label indicating the known or potential association between the mammal and the virus. This label combines information from both the VIRION and GenBank databases to represent whether an association has been documented.
6. **mean_pred**: The mean predicted probability of association across an ensemble of models.



Supplementary Figure SR94 | Predicted increase in associations per mammal species by number of known viral families associations. **a)** Median fold increase in predicted associations versus known associations to viral families (for mammal species with at least one viral family). Log-log scale. **b)** Median probabilities for unlabeled associations versus known associations to viral families. In both the figures, curves show model fits (power law and sigmoid-bounded polynomial, respectively) with 90% confidence intervals.

7. **std_pred**: The standard deviation of a binary presence indicator derived from Monte Carlo simulations over the ensemble predictions. This measures variability in predicted presence/absence classification when applying random thresholds to the predictions.
8. **median_prob**: The median predicted probability of association across the ensemble models, providing a robust central estimate.
9. **std_prob**: The standard deviation of the predicted probabilities across the ensemble.
10. **10quant_prob**: The 10th percentile of predicted probabilities across the ensemble, indicating the lower bound of likely association probability.
11. **90quant_prob**: The 90th percentile of predicted probabilities across the ensemble, indicating the upper bound of likely association probability.
12. **mean_likelihood**: The average likelihood score (i.e., classifier output) for the association across the ensemble models.
13. **median_likelihood**: The median likelihood score across the ensemble, providing a robust central estimate.
14. **std_likelihood**: The standard deviation of the likelihood scores across the ensemble.
15. **10quant_likelihood**: The 10th percentile of likelihood scores across the ensemble.
16. **90quant_likelihood**: The 90th percentile of likelihood scores across the ensemble.
17. **mean_prop**: The mean propensity score for the association.
18. **median_prop**: The median propensity score, providing a robust central estimate.
19. **std_prop**: The standard deviation of the propensity scores.
20. **10quant_prop**: The 10th percentile of propensity scores.
21. **90quant_prop**: The 90th percentile of propensity scores.

Supplementary Notes

1 Random continuous variables in $[0, 1]$

We describe a method for generating continuous random variables constrained within the interval $[0, 1]$ using Beta distributions, ensuring their probability distribution has a mode different from 0 or 1.

Suppose our variable x is described by a Beta distribution:

$$\text{Beta}(x; \alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha + \beta)} x^{\alpha-1} (1-x)^{\beta-1}. \quad (\text{SN1})$$

A Beta distribution has mean and variance given by:

$$\mu(\alpha, \beta) = \frac{\alpha}{\alpha + \beta}, \quad (\text{SN2})$$

$$\sigma^2(\alpha, \beta) = \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)}. \quad (\text{SN3})$$

If we invert the relations we obtain:

$$\alpha(\mu, \sigma) = \mu \left(\frac{\mu(1-\mu)}{\sigma^2} - 1 \right), \quad (\text{SN4})$$

$$\beta(\mu, \sigma) = (1-\mu) \left(\frac{\mu(1-\mu)}{\sigma^2} - 1 \right). \quad (\text{SN5})$$

In particular, since we have $\alpha, \beta > 0$, we obtain that $\mu \in (0, 1)$ and $\sigma^2 \in (0, \mu(1-\mu))$.

However, with α or β lower than 1, the Beta distribution's probability density function diverges for $x \rightarrow 0$ or $x \rightarrow 1$, respectively. This is not a behavior we are expecting for, as we want the mode of the distribution to be different from 0 or 1. Thus, we impose an additional condition on σ^2 , such that both $\alpha, \beta \geq 1$, which leads us to:

$$0 < \sigma^2 \leq \min \left\{ \mu^2 \frac{1-\mu}{1+\mu}, \mu \frac{(1-\mu)^2}{2-\mu} \right\}. \quad (\text{SN6})$$

Thus we can think of σ^2 to be a deterministic function of μ and another stochastic variable $\nu \in (0, 1]$, which we refer to as the *variability scaling factor* of the distribution. Specifically,

$$\sigma^2(\mu, \nu) = \nu \times \min \left\{ \mu^2 \frac{1-\mu}{1+\mu}, \mu \frac{(1-\mu)^2}{2-\mu} \right\}. \quad (\text{SN7})$$

Finally, we can think of α and β to be deterministic function of μ and ν , which can instead be viewed as the parameters defining the probability distribution of x , i.e.,

$$x|\mu, \nu \sim \text{Beta}(x; \alpha(\mu, \nu), \beta(\mu, \nu)). \quad (\text{SN8})$$

In particular, we notice that for $\mu = 0.5$ and $\nu \rightarrow 1$, we recover $\alpha = 1$ and $\beta = 1$, which corresponds to the uniform distribution between 0 and 1. Conversely, for μ fixed and $\nu \rightarrow 0$, the Beta distribution converges to a Dirac delta function $\delta(\mu)$. Thus, although these cases form a null measure set, by allowing μ and ν to be sampled by distributions supported on $(0, 1)$, we still consider both the scenario where the x variables are practically independent and the case where they are practically constrained to be identical. However, we still exclude scenarios where the variance of the Beta distribution in Supplementary Equation SN8 is so high that the most likely values are 0 or 1, and where the distribution becomes bimodal.

2 Positive-Unlabeled measures

In Positive-Unlabeled (PU) scenarios, it is not obvious how to compute standard evaluation metrics because true labels are known only for a subset of the data, and the negative class is not explicitly identified. This introduces challenges for both model evaluation and hyperparameter tuning.

2.1 Metrics for PU learning

The most commonly used metric for tuning using PU data is based on the F_1 score, which is defined as the harmonic mean of the precision and recall, that is:

$$F_1 = \frac{2pr}{p+r}, \quad (\text{SN9})$$

where $p = \Pr(\hat{y} = 1|y = 1)$ is the precision, and $r = \Pr(y = 1|\hat{y} = 1)$ is the recall. In order to get a high F_1 score, both precision and recall must be high. Unfortunately we do not know how to estimate the F_1 score from positive and unlabeled examples. However, a similar score can be evaluated. The following performance criterion, proposed by [46], shares the F_1 score's property of being high when both precision and recall are high:

$$\begin{aligned} \text{PU}F_1 &= \frac{pr}{\Pr(y = 1)} = \frac{pr^2}{r \Pr(y = 1)} = \\ &= \frac{\Pr(y = 1|\hat{y} = 1)r^2}{\Pr(\hat{y} = 1 \wedge y = 1)} = \frac{r^2}{\Pr(\hat{y} = 1)}. \end{aligned} \quad (\text{SN10})$$

Under the SCAR assumption, recall can be consistently estimated from PU data as $\hat{r} = \Pr(y = 1|s = 1)$, enabling the evaluation of the $\text{PU}F_1$ metrics. However, an analogue for the SAR scenario has still not been proposed in literature. We notice that the $\text{PU}F_1$ metric is not bounded from above. Also, while it is proportional to the product of recall and precision, the constant of proportionality, $1/\Pr(y = 1)$, is typically unknown and can vary depending on the dataset and problem. Therefore, this metric can help identify results that maximize the recall-precision product within the same dataset, but $\text{PU}F_1$ scores are not directly comparable across different datasets and problems.

2.2 Extending $\text{PU}F_1$ to SAR scenario

Here, given an instance space Ω , we want to find a way to estimate, from a PU dataset under the SAR scenario, the recall for a model $g : \Omega \rightarrow [0, 1]$ with model's parameters θ , such that, $g : x \mapsto \Pr(y = 1|x, \theta)$. We notice that it is possible to rewrite the recall as:

$$\begin{aligned} r &= \Pr(\hat{y} = 1|y = 1, \theta) = \sum_{x \in \Omega} \Pr(\hat{y} = 1, x|y = 1, \theta) = \\ &= \sum_{x \in \Omega} \Pr(\hat{y} = 1|y = 1, x, \theta) \Pr(x|y = 1, \theta) = \\ &= \sum_{x \in \Omega} \mathbb{1}(g(x) \geq t) \Pr(x|y = 1) = \\ &= \sum_{x \in \Omega} \mathbb{1}(g(x) \geq t) \frac{\pi_s/\pi_y}{e(x)} \Pr(x|s = 1), \end{aligned} \quad (\text{SN11})$$

where we have used the fact $\Pr(\hat{y} = 1|y = 1, x, \theta) = \Pr(\hat{y} = 1|x, \theta) = \mathbb{1}(g(x) \geq t)$ and $\Pr(x|y = 1, \theta) = \Pr(x|y = 1)$, and:

$$\begin{aligned} \Pr(x|s = 1) &= \Pr(x|s = 1, y = 1) = \frac{\Pr(s = 1|x, y = 1) \Pr(x|y = 1)}{\Pr(s = 1|y = 1)} \\ &= \frac{\Pr(s = 1|x, y = 1) \Pr(x|y = 1)}{\Pr(y = 1|s = 1)(\pi_s/\pi_y)} = \frac{e(x)}{\pi_s/\pi_y} \Pr(x|y = 1), \end{aligned} \quad (\text{SN12})$$

with $\pi_s = \Pr(s = 1)$ and $\pi_y = \Pr(y = 1)$. Here, $t \in (0, 1)$ represents the threshold value after which an instance is classified as positive, typically set to 0.5. In particular, in the SCAR scenario the propensity score is independent of the instance x considered, i.e., $e(x) \equiv \Pr(s = 1|y = 1) = \pi_s/\pi_y$, and the relation $\Pr(\hat{y} = 1|y = 1, \theta) = \Pr(\hat{y} = 1|s = 1, \theta)$ is recovered.

That been said, we can rewrite the $\text{PU}F_1$ for SAR scenario as:

$$\text{PU}F_1 = \frac{\pi_s^2}{\pi_y^2} \frac{1}{\sum_{x \in \Omega} \mathbb{1}(g(x) \geq t)} \left[\sum_{x \in \Omega} \frac{\mathbb{1}(g(x) \geq t)}{e(x)} p(x|s = 1) \right]^2. \quad (\text{SN13})$$

However, since the class prior is not known, this score cannot generally be evaluated. Nonetheless, because both the class prior (π_y) and label prior (π_s) are constant for the same dataset and problem, we can omit them in the calculation and define our SAR- $\text{PU}F_1$ score as:

$$\text{SAR-PU}F_1 = \frac{1}{\sum_{x \in \Omega} \mathbb{1}(g(x) \geq t)} \left[\sum_{x \in \Omega} \frac{\mathbb{1}(g(x) \geq t)}{e(x)} p(x|s = 1) \right]^2 \propto pr. \quad (\text{SN14})$$

Similarly, Supplementary Equation SN11 gives us an operative way for estimating recall in a PU scenario respecting the SAR assumption. In fact, we can define:

$$\text{SAR-PU}r = \sum_{x \in \Omega} \frac{\mathbb{1}(g(x) \geq t)}{e(x)} \Pr(x|s=1) \propto r, \quad (\text{SN15})$$

which we know to be proportional to the true recall via the proportionality constant π_s/π_y , which is characteristic of the dataset and problem considered.

Finally, taking the ratio of $\text{SAR-PU}F_1$ and $\text{SAR-PU}r$, we can obtain a metric which is proportional to precision in PU scenarios:

$$\text{SAR-PU}p = \frac{1}{\sum_{x \in \Omega} \mathbb{1}(g(x) \geq t)} \left[\sum_{x \in \Omega} \frac{\mathbb{1}(g(x) \geq t)}{e(x)} p(x|s=1) \right] \propto p. \quad (\text{SN16})$$

$\text{SAR-PU}r$, $\text{SAR-PU}F_1$, and $\text{SAR-PU}p$ cannot provide information about a model's absolute performance, as they are only proportional to traditional metrics through unknown proportionality constants. However, they serve as reliable tools for comparing different models on the same dataset.

In particular, in SCAR scenarios, these metrics from different models can be easily confronted. In SAR ones, precise estimations of the propensity scores $e(x)$ are required. This complicates confronting results of different works on the same dataset, as propensity scores may vary depending on how they are evaluated.

2.3 Naive evaluation metrics in PU scenarios

Standard metrics, obtained treating labeled instances as positive and unlabeled instances as negative, usually fail to describe the actual performance of a classifier in a PU scenario. However, they can still give us some insights on the quality of the predictions.

Naive-Recall and Naive-NegRecall

We define the Naive-Recall as:

$$\text{Naive-Recall} = \frac{\#\text{labeled samples s.t. } \text{pred}(m, v) > 0.5}{\#\text{labeled samples}}, \quad (\text{SN17})$$

and the Naive-NegRecall as:

$$\text{Naive-NegRecall} = \frac{\#\text{unlabeled samples s.t. } \text{pred}(m, v) < 0.5}{\#\text{unlabeled samples}}. \quad (\text{SN18})$$

Naive-Recall quantifies the ability of the model to recover labeled instances as positive, while Naive-NegRecall the tendency to predict unlabeled samples as negative.

A high Naive-Recall is generally desirable, as it indicates correct identification of known positives. Importantly, in a SCAR scenario, Naive-Recall serves as an unbiased estimator of the true recall $r = \Pr(\hat{y} = 1|y = 1)$. Generally, when propensity scores are approximately uniform across instances, Naive-Recall remains a reliable proxy for true recall.

In contrast, the interpretation of Naive-NegRecall depends on how many true positives are expected among the unlabeled set. Still, it indicates how conservative predictions are.

Naive-Precision

We define the Naive-Precision as:

$$\text{Naive-Precision} = \frac{\#\text{labeled samples s.t. } \text{pred}(m, v) > 0.5}{\#\text{samples s.t. } \text{pred}(m, v) > 0.5}. \quad (\text{SN19})$$

Models with higher Naive-Precision are more conservative, whereas models with lower Naive-Precision tend to produce predictions that deviate more significantly from the observations.

Naive-ROC and Naive-PRAUC

We define Naive-ROC and Naive-PRAUC scores in analogy to standard ROC and PRAUC metrics: labeled instances are treated as positives, while unlabeled instances as negatives. These scores reflect how well the model distinguishes labeled from unlabeled data. However, since unlabeled data includes both positive and negative instances, using these metrics to evaluate models' preferences requires caution.

The foundation of PU learning is to recover unbiased predictions for both labeled and unlabeled instances. For this reason, a key property of PU learning is that classifiers are not always penalized for assigning high probabilities to unlabeled instances. As a result, PU-trained classifiers are not explicitly incentivized to assigning higher probabilities to labeled instances than to unlabeled ones, which undermines the optimal ranking assumption underlying ROC and PRAUC metrics.

In fact, models trained by naively treating unlabeled instances as negatives often yield higher Naive-ROC and Naive-PRAUC scores than those models that are properly trained PU models, even when the latter more accurately capture the true distribution of positives. Conservative models in particular tend to perform better under these naive metrics. Nevertheless, because labeled data consists only of true positives and unlabeled data includes a mix of positives and negatives, we do expect, on average, labeled instances to be predicted with higher probabilities. Therefore, with appropriate caution, Naive-ROC and Naive-PRAUC can still provide useful insights into models' performances in PU settings.

3 Pseudocode of the algorithm for partitioning associations

Here, we give the reader with the algorithm used for partitioning associations in a given set of mammal-virus associations in K balanced and non-overlapping subsets.

Algorithm 2: Splitting associations into K balanced folds

```

Input : associations to be split  $\mathcal{E}$ 
         number of folds  $K$ 
Output: associations split into folds  $(\mathcal{E}_1, \dots, \mathcal{E}_K)$ 

1 for  $k = 1$  to  $K$  do
2    $\mathcal{E}_k \leftarrow$  empty set;

3  $\text{viral\_families} \leftarrow \text{Shuffle}(\text{viral\_families});$ 
4 foreach  $v_0 \in \text{viral\_families}$  do
5    $\text{mammalian\_degrees} \leftarrow \text{Shuffle}(\text{mammalian\_degrees});$ 
6   foreach  $\text{degree} \in \text{mammalian\_degrees}$  do
7      $\text{mammalian\_orders} \leftarrow \text{Shuffle}(\text{mammalian\_orders});$ 
8     foreach  $\text{order} \in \text{mammalian\_orders}$  do
9        $\text{mammalian\_families} \leftarrow \text{Shuffle}(\text{mammalian\_families});$ 
10      foreach  $\text{family} \in \text{mammalian\_families}$  do
11         $\mathcal{E} \leftarrow \text{Shuffle}(\mathcal{E});$ 
12        foreach  $(m, v) \in \mathcal{E}$  do
13          if  $v = v_0 \wedge \text{degree}(m) = \text{degree} \wedge$ 
14             $\text{order}(m) = \text{order} \wedge \text{family}(m) = \text{family}$  then
15            for  $k = 1$  to  $K$  do
16               $\alpha_k \leftarrow |\mathcal{E}_k| \times 10^{20};$ 
17              foreach  $(m', v') \in \mathcal{E}_k$  do
18                if  $v' = v_0$  then
19                   $\alpha_k \leftarrow \alpha_k + 10^{16};$ 
20                if  $\text{degree}(m') = \text{degree}$  then
21                   $\alpha_k \leftarrow \alpha_k + 10^{12};$ 
22                if  $\text{order}(m') = \text{order}$  then
23                   $\alpha_k \leftarrow \alpha_k + 10^8;$ 
24                if  $\text{family}(m') = \text{family}$  then
25                   $\alpha_k \leftarrow \alpha_k + 10^4;$ 
26                if  $m' = m$  then
27                   $\alpha_k \leftarrow \alpha_k + 1;$ 
28               $\text{indexes} \leftarrow \text{argmin}(\{\alpha_k\}_{k=1}^K);$ 
29               $k_0 \leftarrow \text{RandomSelect}(\text{indexes});$ 
30               $\mathcal{E}_{k_0} \leftarrow \mathcal{E}_{k_0} \cup \{(m, v)\};$ 

31 return  $(\mathcal{E}_1, \dots, \mathcal{E}_K)$ 

```

Supplementary Information References

1. IUCN. *The IUCN Red List of Threatened Species. Version 2020-2* 2020. <https://www.iucnredlist.org>.
2. Tonelli, A., Blagrove, M., Wardeh, M. & Di Marco, M. *A framework to predict zoonotic reservoirs under data uncertainty: a case study on betacoronaviruses* Apr. 2024.
3. Albery, G. F. & Becker, D. J. Fast-lived Hosts and Zoonotic Risk. *Trends in Parasitology* **37**, 117–129 (2021).
4. Soria, C. D., Pacifici, M., Di Marco, M., Stephen, S. M. & Rondinini, C. COMBINE: a coalesced mammal database of intrinsic and extrinsic traits. *Ecology* **102**, e03344 (2021).
5. Faurby, S. *et al.* *PHYLACINE 1.2.1: An update to the Phylogenetic Atlas of Mammal Macroecology* 2020.
6. Dwivedi, V. P. & Bresson, X. *A Generalization of Transformer Networks to Graphs* 2021. arXiv: 2012.09699 [cs.LG].
7. Downs, C. J., Doctermann, N. A., Ball, R., Klasing, K. C. & Martin, L. B. The Effects of Body Mass on Immune Cell Concentrations of Mammals. *The American Naturalist* **195**, 107–114 (2020).
8. Sheldon, B. C. & Verhulst, S. Ecological immunology: costly parasite defences and trade-offs in evolutionary ecology. *Trends in Ecology & Evolution* **11**, 317–321 (1996).
9. Villemereuil, P. d., Wells, K., Edwards, R. D. & Mouquet, N. Bayesian models for comparative analysis integrating phylogenetic uncertainty. *BMC Evolutionary Biology* **12**, 102 (2012).
10. Faurby, S. & Svenning, J.-C. A species-level phylogeny of all extant and late Quaternary extinct mammals using a novel heuristic-hierarchical Bayesian approach. *Molecular Phylogenetics and Evolution* **84**, 14–26 (2015).
11. Huang, S., Bininda-Emonds, O. R. P., Stephens, P. R., Gittleman, J. L. & Altizer, S. Phylogenetically related and ecologically similar carnivores harbour similar parasite assemblages. *Journal of Animal Ecology* **83**, 671–680 (2014).
12. Wells, K. *et al.* Global spread of helminth parasites at the human-domestic animal-wildlife interface. *Global Change Biology* **24**, 3254–3265 (2018).
13. Stephens, P. R. *et al.* Parasite sharing in wild ungulates and their predators: effects of phylogeny, range overlap, and trophic links. *Journal of Animal Ecology* **88**, 1017–1028 (2019).
14. Cooper, N. *et al.* Phylogenetic host specificity and understanding parasite sharing in primates. *Ecology Letters* **15**, 1370–1377 (2012).
15. Albery, G. F., Eskew, E. A., Ross, N. & Olival, K. J. Predicting the global mammalian viral sharing network using phylogeography. *Nature Communications* **11**, 2260 (2020).
16. Wardeh, M., Blagrove, M. S. C., Sharkey, K. J. & Baylis, M. Divide-and-conquer: machine-learning integrates mammalian and viral traits with network features to predict virus-mammal associations. *Nature Communications* (July 2021).
17. Sanjuán, R., Nebot, M. R., Chirico, N., Mansky, L. M. & Belshaw, R. Viral Mutation Rates. *Journal of Virology* **84** (2010).
18. Mahy, B. W. The Evolution and Emergence of RNA Viruses. *Emerging Infectious Diseases* **16**, 899 (2010).
19. Pybus, O. G., Tatem, A. J. & Lemey, P. Virus evolution and transmission in an ever more connected world. *Proceedings of the Royal Society B: Biological Sciences* **282** (2015).
20. Coffin, J. M. in *The Retroviridae* 19–49 (Springer US, 1992).
21. Nisole, S. & Saïb, A. Early steps of retrovirus replicative cycle. *Retrovirology* **1** (2004).
22. Wawrzyniak, P., Plucienniczak, G. & Bartosik, D. The different faces of rolling-circle replication and its multi-functional initiator proteins. *Frontiers in Microbiology* **8** (2017).
23. Lin, X. *et al.* Order and disorder control the functional rearrangement of influenza hemagglutinin. *Proceedings of the National Academy of Sciences of the United States of America* **111**, 12049–12054 (2014).
24. Lowen, A. C. Constraints, Drivers, and Implications of Influenza A Virus Reassortment. *Annual Review of Virology* **4**, 105–121 (2017).

25. Rey, F. A. & Lok, S.-M. Common Features of Enveloped Viruses and Implications for Immunogen Design for Next-Generation Vaccines. *Cell* **172**, 1319–1334 (2018).
26. Allen, T., Murray, K. A., Zambrana-Torrel, C., *et al.* Global hotspots and correlates of emerging zoonotic diseases. *Nature Communications* **8**, 1124 (2017).
27. Becker, D. J., Crowley, D. E., Washburne, A. D. & Plowright, R. K. Temporal and spatial limitations in global surveillance for bat filoviruses and henipaviruses. *Biology Letters* **15**, 20190423. <https://doi.org/10.1098/rsbl.2019.0423> (2019).
28. Dutilh, B. E., Reyes, A., Hall, R. J. & Whiteson, K. L. Editorial: Virus discovery by metagenomics: the (Im)possibilities. *Frontiers in Microbiology* **8**, 1710 (2017).
29. Van Dam, A., Dekker, M., Morales-Castilla, I., *et al.* Correspondence analysis, spectral clustering and graph embedding: applications to ecology and economic complexity. *Scientific Reports* **11**, 8926 (2021).
30. Udvardy, M. D. F. *A classification of the biogeographical provinces of the world* IUCN Occasional Paper no. 18 (IUCN, Morges, Switzerland, 1975).
31. Bekker, J. & Davis, J. Learning from positive and unlabeled data: a survey. *Mach Learn* **109**, 719–760 (2020).
32. Plummer, M. *JAGS: A Program for Analysis of Bayesian Graphical Models Using Gibbs Sampling* in *Proceedings of the 3rd International Workshop on Distributed Statistical Computing (DSC 2003)* (Vienna, Austria, Mar. 2003). <http://mcmc-jags.sourceforge.net/>.
33. Foundation, P. S. *Python: A Comprehensive Guide to the Python Programming Language* <https://www.python.org/> (2023). <https://www.python.org/>.
34. Vehtari, A., Gelman, A., Simpson, D., Carpenter, B. & Bürkner, P.-C. Rank-Normalization, Folding, and Localization: An Improved \hat{R} for Assessing Convergence of MCMC (with Discussion). *Bayesian Analysis* **16**, 667–718 (2021).
35. Bompiani, E., Petrillo, U. F., Jona Lasinio, G. & Palini, F. *High-Performance Computing with TeraStat* in *2020 IEEE Intl Conf on Dependable, Autonomic and Secure Computing, Intl Conf on Pervasive Intelligence and Computing, Intl Conf on Cloud and Big Data Computing, Intl Conf on Cyber Science and Technology Congress (DASC/PiCom/CBDCom/CyberSciTech)* (2020), 499–506.
36. Kipf, T. N. & Welling, M. *Semi-Supervised Classification with Graph Convolutional Networks* in *International Conference on Learning Representations* (2017).
37. Yun, S., Jeong, M., Kim, R., Kang, J. & Kim, H. J. *Graph Transformer Networks* in *Advances in Neural Information Processing Systems* **32** (Curran Associates, Inc., 2019).
38. He, K., Zhang, X., Ren, S. & Sun, J. Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification. *arXiv preprint arXiv:1502.01852*. [cs.CV] (2015).
39. Li, G., Xiong, C., Thabet, A. & Ghanem, B. *DeeperGCN: All You Need to Train Deeper GCNs* 2020. arXiv: 2006.07739 [cs.LG].
40. Bekker, J., Robberechts, P. & Davis, J. *Beyond the Selected Completely at Random Assumption for Learning from Positive and Unlabeled Data* in *Machine Learning and Knowledge Discovery in Databases* (Springer International Publishing, 2020), 71–85.
41. Kiryo, R., Niu, G., du Plessis, M. C. & Sugiyama, M. *Positive-Unlabeled Learning with Non-Negative Risk Estimator* 2017. arXiv: 1703.00593 [cs.LG].
42. Loshchilov, I. & Hutter, F. Decoupled Weight Decay Regularization. *arXiv preprint arXiv:1711.05101* (2017).
43. Kingma, D. P. & Ba, J. Adam: A Method for Stochastic Optimization. *arXiv preprint arXiv:1412.6980* (2014).
44. Micikevicius, P. *et al.* Mixed Precision Training. *arXiv preprint arXiv:1710.03740* (2017).
45. Saito, Y., Yaginuma, S., Nishino, Y., Sakata, H. & Nakata, K. *Unbiased Recommender Learning from Missing-Not-At-Random Implicit Feedback* in *Proceedings of the 13th International Conference on Web Search and Data Mining* (Association for Computing Machinery, New York, NY, USA, 2020), 501–509.
46. Lee, W. S. & Liu, B. *Learning with Positive and Unlabeled Examples Using Weighted Logistic Regression* in *Proceedings of the Twentieth International Conference on Machine Learning* (2003), 448–455.