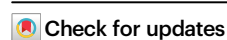


Predicting unknown mammalian viral hosts with Dynamic Positive–Unlabeled learning

Received: 14 October 2025

Accepted: 24 July 2026

Published online: 17 September 2026

Gabriele Pignalberi ¹✉, Andrea Tonelli ², Stefano Giagu ¹ & Moreno Di Marco ²✉

Most emerging infectious diseases in humans are zoonotic in origin. Yet, knowledge of host–pathogen associations remains limited. Machine learning models have been increasingly used to infer missing zoonotic host–pathogen associations, but face challenges due to biased data and the absence of confirmed negative cases. Here, we introduce Dynamic Positive–Unlabeled (DPU) learning, a semi-supervised framework for predicting missing links in incomplete, biased networks. DPU learning integrates a propensity score model that estimates the probability of observing existing links with a classifier that predicts true link existence. This approach corrects predictions for uneven sampling bias and explicitly accounts for missing links arising from either true absences of associations or gaps in data collection. We applied DPU learning to predict missing associations between 5,330 mammalian species and 33 viral families worldwide, integrating phylogeographic relationships between mammals, observed mammal–virus association patterns, mammalian traits, and viral genetic characteristics. Our model predicted newly confirmed associations absent from the training data with a median probability of 0.82 and estimated a 8.6-fold increase in the total number of mammalian species–viral family associations. These results highlight the potential of DPU learning to support strategic, data-driven surveillance activities for proactive zoonotic risk mitigation.

Two-thirds of emerging infectious diseases have zoonotic origin, with wild animals being the most common reservoirs of pathogens^{1,2}. Mammals alone are predicted to harbor hundreds of thousands of undiscovered viruses^{3,4}, with highly divergent estimates of how many of these have zoonotic potential^{3–5}. Humans share several viruses with other mammals, e.g., HIV, Rabies virus, Ebola virus, SARS-CoV-2, some of which have had profound consequences for public health and human livelihoods⁶. Preventing the spillover of zoonotic viruses into humans would be orders of magnitude cheaper than facing the consequences of a disease outbreak⁶, but this requires a comprehensive knowledge of pathogens existing in the wild. With only a small fraction of the global mammal virome currently described^{3,4,7,8}, knowledge of existing pathogens is far from comprehensive, and establishing

effective pandemic preparedness strategies remains a global challenge⁹.

The enormous knowledge gap on the diversity of mammalian viruses⁴ requires accelerated discovery rates to be filled³. At the moment, however, resources are largely insufficient to achieve this acceleration, calling for the development of data-driven methodologies to support traditional virological sampling strategies¹⁰. Predictive modeling has been employed to aid the identification of potential host species to be targeted for zoonotic risk surveillance^{11,12}. Such approaches leverage the patterns that characterize the known host–virus associations, correlating observed susceptibilities with biological and ecological determinants of viral exposure and infection, to predict the existence of unknown associations^{13,14}. The growth of machine learning

¹Department of Physics, Sapienza University of Rome, Rome, Italy. ²Department of Biology and Biotechnologies ‘Charles Darwin’, Sapienza University of Rome, Rome, Italy. ✉e-mail: gabriele.pignalberi@uniroma1.it; moreno.dimarco@uniroma1.it

approaches has driven rapid development in the predictions of host–virus associations^{15,16}, but, despite methodological developments, important challenges remain unaddressed.

First, observed mammal–virus associations are heavily influenced by sampling biases, as research efforts have disproportionately targeted certain taxa and geographic areas following public health emergencies and logistical convenience^{17,18}. Second, the absence of confirmed true negatives further complicates the training of machine learning predictive models: a single recorded occurrence of a virus in a mammal is usually deemed sufficient as evidence of an existing association (i.e., positive case), but the lack of records cannot be taken as evidence of a nonexistent association (i.e., negative case). As a result, predicting missing associations should be cast as a positive–unlabeled (PU) learning problem, where only a fraction of the positive associations is labeled, and the rest—potentially containing both positives and negatives—remains unlabeled.

Under this PU framework, it is crucial to distinguish two fundamental mechanisms that underlie data generation. The first is the ground truth mechanism, which can be modelled through a classifier that estimates the probability that candidate associations truly exist. The second is the labeling mechanism, which can be modelled through a propensity score model that estimates the probability that existing associations are empirically observed and thus labeled. By jointly modeling these two complementary mechanisms, it is possible to effectively disentangle real biological patterns from observational biases and improving models' fidelity¹⁹.

Although previous studies have recognized the challenges posed by sampling bias and the absence of true negatives in host–pathogen networks^{20–22}, most practical implementations have treated unlabeled associations as negatives^{11,23–26}. This effectively imposes a naïve closed-world assumption²⁷, whereby all undocumented associations—including those that are simply unobserved—are assumed not to exist. Failing to explicitly account for the PU nature of the data can result in unwarranted and biased predictions that may reflect observational artifacts rather than true associations²⁸, a problem worsened in the presence of uneven sampling efforts. Some methodological efforts have attempted to relax the closed-world assumption. These include approaches that correct for class imbalance between positive and unlabeled cases (e.g., through synthetic oversampling of known positives²⁴) or that designate subsets of unlabeled associations as pseudo-negatives based on their probability of being observed²¹. While such strategies represent important pragmatic improvements over a naïve closed-world assumption, they still ultimately reduce the problem to a conventional positive-negative classification task.

Alternative approaches have adopted a Bayesian perspective, explicitly separating the ground-truth from the labeling processes and thereby providing a principled solution to the conceptual limitations of the closed-world assumption. However, Bayesian formulations to host–pathogen networks have typically relied on uniform propensity scores across all possible associations^{22,29}, equivalent to the Selected Completely at Random (SCAR) assumption³⁰, wherein the probability of observing an existing association is independent of the instance under consideration. This simplification fails to account for the substantial heterogeneity in sampling effort across host and pathogen taxa, as well as geographic regions. Hierarchical Bayesian models have been introduced in the ecological network literature to accommodate instance-specific detectability conditional on true process existence^{31,32}, consistent with the Selected At Random (SAR) assumption³³, under which propensity scores depend on observed covariates. To our knowledge, however, such hierarchical Bayesian approaches have not been extended to the prediction of missing associations in host–pathogen networks. Furthermore, while Bayesian methods can provide principled uncertainty quantification, they remain limited in their capacity to capture complex, high-dimensional, and graph-structured data patterns—a domain in which recent

advances in representation learning have been shown to be highly effective^{34–36}.

Furthermore, while both SAR and SCAR assumptions treat propensity scores as static and independent of previously labeled instances, the probability that an existing host–pathogen association is observed increases as knowledge accumulates. Each new study raises the likelihood that true but undocumented associations will be detected, and newly discovered associations frequently redirect sampling effort toward related geographic areas or phylogenetically proximate host and pathogen taxa. Consequently, propensity scores are more accurately characterized as intrinsically dynamic, monotonically non-decreasing functions of time, shaped by historical sampling activities and previously documented associations¹⁸. This motivates a propensity score model that explicitly accounts for the evolving set of observed associations, incorporating also contextual information derived from currently documented associations. Likewise, because newly documented associations increasingly reveal informative biological structure, classifiers can also benefit from conditioning their predictions on the evolving observed network rather than relying solely on static covariates or prefixed relational or hierarchical structures.

Here, we introduce Dynamic Positive–Unlabeled (DPU) learning (Fig. 1), a novel semi-supervised framework for building quantitative machine learning predictive models in PU scenarios where predictions are explicitly conditioned on currently labeled data. This formulation allows the classifier to infer biologically plausible missing links based on the current structure of the observed host–pathogen network. Similarly, the propensity score model can use the same information to characterize sampling biases. We apply the DPU framework to the network of observed mammal–virus associations to predict missing links between 5,330 terrestrial mammal species and 33 viral families. In this context, a predicted association between a mammal species and a viral family may suggest either a known virus that has not yet been sampled in that host or a novel virus from that family infecting the host. Contrary to prior studies that limited their analysis only to mammals with known viral associations (ca. 27% of all mammal species^{37,38}), we include all terrestrial mammals, providing a more comprehensive picture of global mammal biodiversity and allowing predictions also for hosts without known viral association. We adopt a hierarchical, network-informed Bayesian model for estimating propensity scores across mammal–virus pairs, allowing us to explicitly quantify heterogeneity in research effort and correcting for imperfect and uneven detection. Using these propensity scores, we then train an ensemble of graph neural networks (GNNs) to infer biologically plausible associations, leveraging mammal and virus traits, host phylogeography, and emerging patterns in the observed mammal–virus network. By explicitly correcting for observation biases and incomplete sampling, DPU learning overcomes some of the customary challenges in predicting unknown viral hosts, delivering improved predictions to guide viral surveillance strategies.

Results

Performance of DPU learning in predicting missing mammal–virus associations

We assessed the predictive performance and robustness of our DPU framework for mammal–virus associations using a 10-fold cross-validation (see the “Methods” section). All mammal–virus pairs (both labeled and unlabeled) were partitioned into 10 non-overlapping folds. In each iteration, eight folds were used for training, one for validation, and one as a held-out test set. Mean model performance was stable across folds (Supplementary Table S1A): training sets had a mean naïve recall of 0.99, naïve ROC-AUC of 0.95, and naïve PR-AUC of 0.31; validation sets had a mean naïve recall of 0.90, naïve ROC-AUC of 0.91, and naïve PR-AUC of 0.27; test sets had a mean naïve recall of 0.88, naïve ROC-AUC of 0.91, and naïve PR-AUC of 0.26. Here, naïve denotes

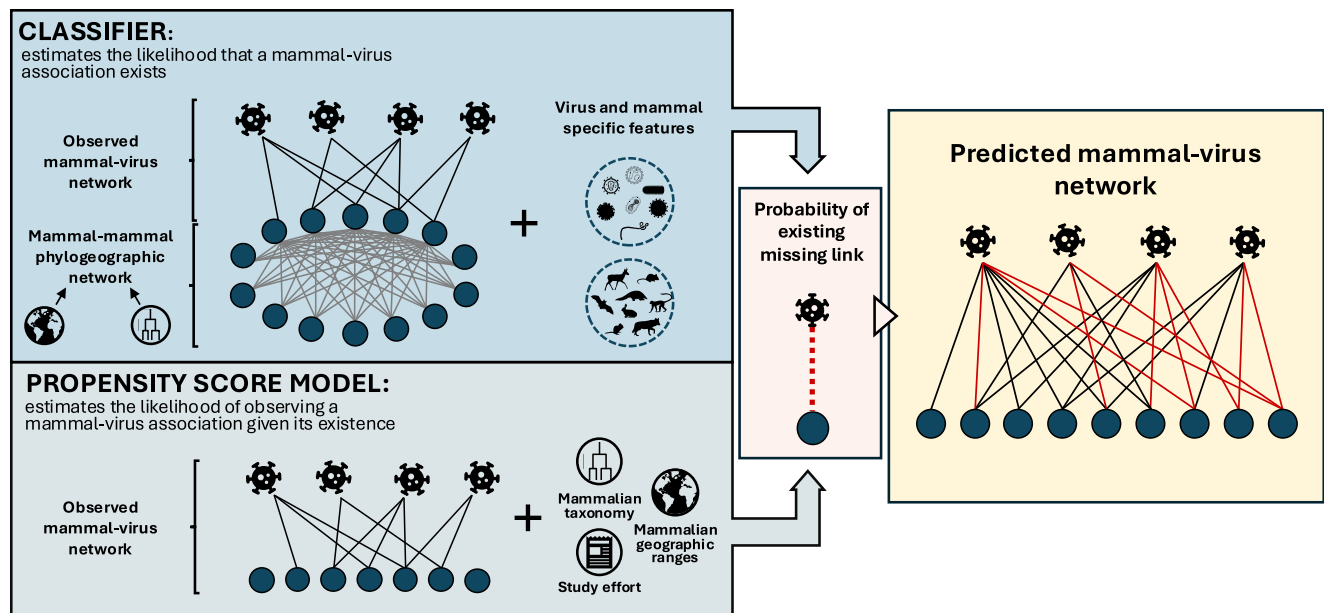


Fig. 1 | Dynamic Positive-Unlabeled (DPU) framework for predicting missing mammal-virus associations. The DPU framework combines a classifier and a propensity score model to jointly account for both the ground truth mechanism, responsible for determining truly existent associations from nonexistent ones, and the labeling mechanism, responsible for which existent associations are effectively documented through scientific evidence. Specifically, the classifier employs a graph neural network (GNN) framework to capture emerging patterns within the observed network of mammal-virus associations, as well as in the mammalian phylogeographic network. Combining these relational structures with both viral-specific and mammal-specific features, the GNN generates node representations that encode both the positional context of the node within the observed network of

mammal-virus interactions and node-specific biological features, enabling the identification of biologically realistic associations between mammal species and viruses. The propensity score model is built using a hierarchical Bayesian framework that incorporates mammalian taxonomy, proxies for study effort, and hosts geographic range data. By integrating these complementary factors, the model provides informed propensity score estimates for all mammal-virus associations, accounting for human-centric, geographic, biological, and self-reinforcing biases. Finally, the classifier and propensity score model are jointly used to estimate the probability that a missing association reflects a true but undocumented link, forming the basis for reconstructing the predicted mammal-virus association network.

that unlabeled associations were treated naively as negatives for metric computation. Although naïve PR-AUC values (0.26–0.31) were modest in absolute terms, they still exceeded the random baseline (–0.03; corresponding to a labeled prevalence of ~2.8%) by approximately an order of magnitude. Moreover, as these metrics treat all unlabeled associations as true negatives, naïve PR-AUC values constitute a conservative lower bound on true model performance. To assess how model performance varied with host sampling intensity, mammal species were grouped by their number of observed viral family associations. Across all groups, median predicted probabilities for labelled test-set associations remained high (medians across groups: mean = 0.993, s.d. = 0.005; Supplementary Fig. S1), indicating consistent recovery of known associations across both well-studied and sparsely sampled species.

We further evaluated the model's ability to identify associations that were treated as unlabeled during our training but were subsequently discovered—referred to hereafter as new labels. This provides an independent test of the model's ability to prioritize biologically real associations among unlabeled candidates, directly emulating the intended deployment scenario in which predictions are made over a set of still-undiscovered associations conditioned on those currently documented. New labels received much higher predicted probabilities in our framework than associations that remained unlabeled (median 0.82 vs. <0.01), confirming the model's discriminatory capacity (Fig. 2a). This pattern held consistently across individual viral families, mammalian orders, and mammal species with varying numbers of known associations (Supplementary Figs. S2–S4). The difference between group-level median predicted probabilities for new labels versus still-unlabeled associations averaged 0.39 (s.d. = 0.28) across viral families, 0.59 (s.d. = 0.33) across mammalian orders, and 0.41 (s.d. = 0.28) across species groups. Discrimination was further

quantified using the still-unlabeled-prevalence-new-label-sensitivity curve (SUPNLSC)²⁵, a novel ROC-style metric evaluating the relative ranking of newly labelled versus still-unlabeled associations. The resulting AUC of 0.81 confirms substantially better-than-random discrimination between new labels and still-unlabeled associations (Fig. 2b).

The DPU framework was further evaluated on two synthetic datasets constructed by augmenting real observational data with synthetic associations belonging to one of two hypothetical viral families. This design afforded full control over the ground truth and labelling mechanisms for the synthetic family while preserving the real-data context for all remaining viruses (see the “Methods” section). For the first synthetic viral family (*Syn1*), all Musteloidea mammals were designated as viable hosts, with subsequent viral spread simulated to Caniformia species inhabiting the Nearctic realm. For the second synthetic family (*Syn2*), all Laurasiatheria mammals in the Afrotropical region were designated as hosts. Predictions from models trained on these datasets aligned well with the ground truth for associations that were provided as unlabeled during training, with true unlabeled synthetic associations consistently ranked higher than non-existent ones (Fig. 2c, e). Predictive performance was robust across both scenarios (*Syn1*: ROC-AUC = 0.90, PR-AUC = 0.60; *Syn2*: ROC-AUC = 0.81, PR-AUC = 0.38; Supplementary Table S2), with the model correctly identifying true host clades, while assigning low predicted probabilities to the majority of non-associated taxa (Supplementary Fig. S5).

The model also demonstrated a notable extrapolative capacity of inference by identifying non-trivial association patterns for taxa with no labeled associations during training. For example, despite the order Pholidota having no labeled associations with *Syn2* during training, the model correctly predicted one of three positive species within this group with high probability (around 0.998). A small number of species

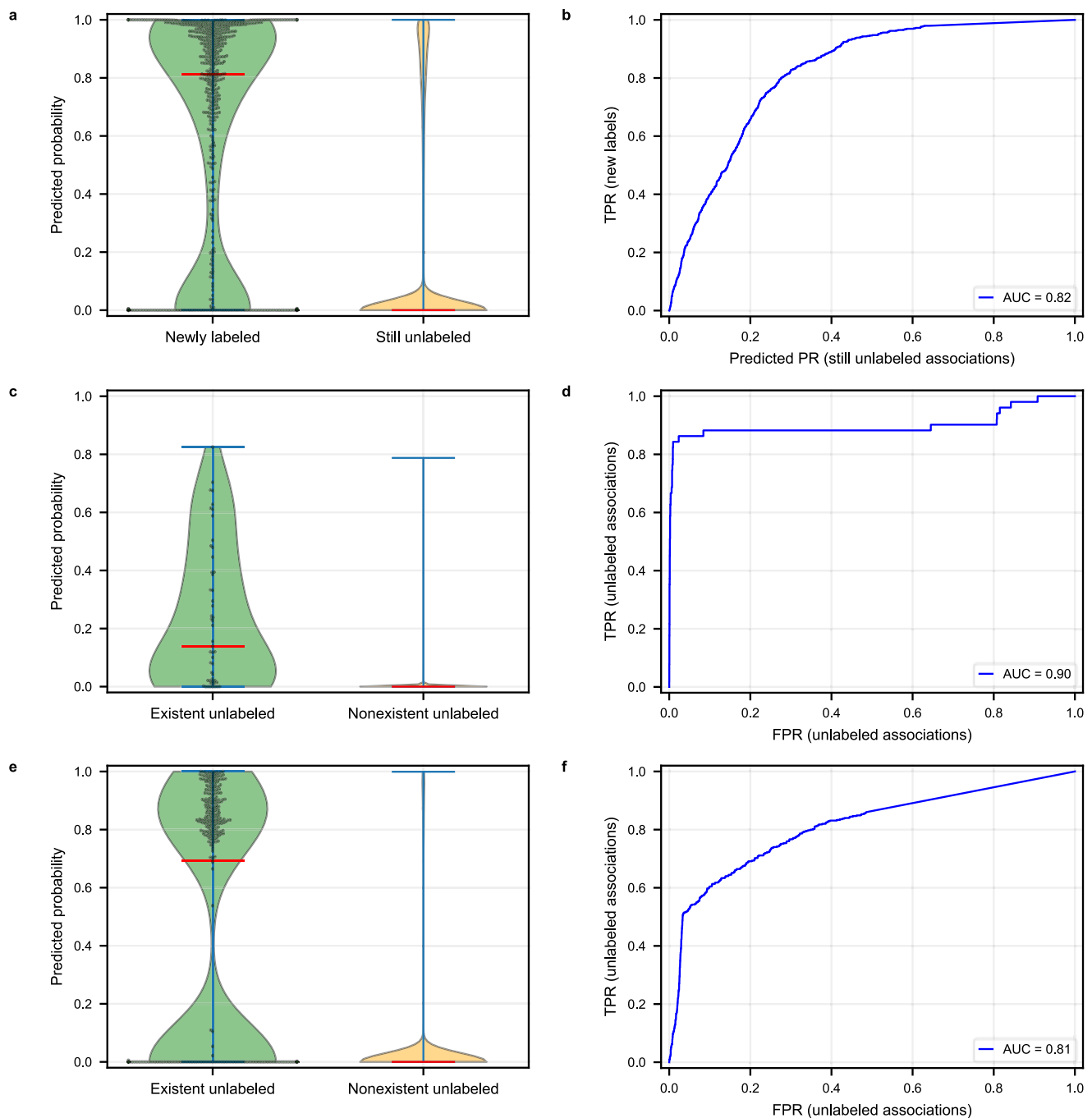


Fig. 2 | Predictions for unlabeled associations using observational and synthetic data. For panels **a**, **c**, **e**: each data point is the predicted probability for one distinct mammal species–viral family pair, averaged across five observed mammal–virus subgraphs (see the “Methods” section); horizontal red lines indicate the median of each distribution, horizontal blue lines indicate the extrema (minimum and maximum) predictions for each group, and violin width reflects the local density of predicted probabilities. Individual data points are not overlaid for groups exceeding 1000 pairs to avoid overplotting. **a** Violin plots comparing predicted probabilities for newly discovered associations (new labels; $n = 686$ pairs) against associations that remained unlabeled ($n = 170,236$ pairs). **b** Still-unlabeled prevalence–new label-sensitivity curve (SUPNLSC), showing the true positive rate

(TPR) of new labels versus the predicted positive rate (PR) for still-unlabeled associations across all thresholds from 0 to 1 (sample sizes as in **a**); area under the curve (AUC) is indicated for the curve. **c**, **e** Violin plots comparing predicted probabilities for true/existent (*Syn1*; $n = 51$ pairs; *Syn2*; $n = 450$ pairs) versus false/nonexistent (*Syn1*; $n = 5249$ pairs; *Syn2*; $n = 4780$ pairs) unlabeled synthetic associations. **d**, **f** Receiver operating characteristic (ROC) curves evaluated on unlabeled synthetic associations for *Syn1* and *Syn2*, respectively (sample sizes as in **c**, **e**), showing the true positive rate (TPR) versus false positive rate (FPR) across unlabeled synthetic associations at each threshold; area under the curve (AUC) is indicated for each curve.

were mispredicted. For instance, the single species in the order Tubulidentata received a high predicted probability despite being a true negative, likely reflecting the higher uncertainty in very small, phylogenetically isolated clades. Other negative species also received relatively high predictions among Primates (25 out of 478 species had

probability >0.05), Hyracoidea (3 out of 5), and Proboscidea (1 out of 2). Although these orders represent true negatives under the ground truth of *Syn2* (falling outside the host clade Laurasiatheria), most predicted species (all except five primates) were still located within the Afrotropical realm. This suggests that, while the model occasionally

over-extended host-clade boundaries, its predictions remained biologically realistic and coherent with the simulated patterns of *Syn2*, successfully capturing the underlying biogeographic constraints defining the true hosts.

For *Syn1*, the model correctly identified a lower degree of geographic constraint, yielding predictions closely aligned with the ground truth distribution even in regions with sparser recorded associations (Supplementary Fig. S6). For *Syn2*, the model appropriately captured a stronger geographic constraint, correctly recovering the Afrotropical hub of viable hosts with few predictions outside this region (Supplementary Fig. S7). ROCs computed for unlabeled synthetic associations against the ground truth showed a steep initial rise before transitioning to a more gradual increase after a well-defined knee corresponding to a near-optimal classification threshold (Fig. 2d, f). This indicates that the models retrieve a subset of true associations with high confidence, producing rapid gains in sensitivity with minimal false positives prior to entering a region of diminishing separability. When using a naïve classification threshold of 0.5, the models achieved high precision but limited recall (*Syn1*: F_1 -score = 0.29, Precision = 0.82, Recall = 0.18; *Syn2*: F_1 -score = 0.544, Precision = 0.59, Recall = 0.51; Table S2). However, optimizing the threshold to maximize the F_1 -score substantially improved sensitivity for *Syn1*, where a threshold of 0.07 yielded a F_1 -score of 0.66 (Precision = 0.70, Recall = 0.63; Table S2). Collectively, these results suggest that the DPU approach is particularly suited to ranking potential hosts, retaining high discriminatory power even when absolute values of predicted probabilities may be low.

Patterns of predicted mammal–virus associations

After validation, we trained an ensemble of 100 DPU models, each conditioned on a different mammal–virus subgraph and configuration of propensity scores drawn from their Bayesian posterior, allowing each model to represent a distinct, plausible scenario of how sampling efforts were distributed and shaped observational data. Across the ensemble, we predicted a median expected number of 42,838 (90% CI = 39,927–47,735) new associations between mammal species and viral families. These predictions correspond to 25.2% (90% CI = 23.5%–28.0%) of all previously unobserved associations at the time of training. As a result, the sparsity of the mammal–virus association matrix decreased from 96.8% to a predicted 72.4% (90% CI = 69.7%–74.1%), representing an 8.6-fold increase (90% CI = 8.1–9.4) in the total number of associations relative to those currently observed. To avoid ambiguity, hereafter, fold increase is defined as the ratio of the total number of predicted associations—including both newly predicted and previously observed true associations—to the number of currently observed associations.

On the host side, both species-poor and speciose mammalian orders showed high numbers of new predicted associations per species. Orders with few living representatives—including Tubulidentata (median 18.4, 90% CI = 1.9–29.2; 1 species), Perissodactyla (median 15.4, 90% CI = 11.3–18.4; 11 species), and Proboscidea (median 14.4, 90% CI = 8.6–19.4; 2 species)—ranked among the highest, but so did speciose orders such as Cetartiodactyla (median 14.5, 90% CI = 13.3–16.1; 213 species) and Primates (median 12.2, 90% CI = 11.2–13.8; 478 species; Supplementary Fig. S8a). In terms of fold increase in the total number of associations, the largest gaps were identified in Afrosoricida (median 52.4, 90% CI = 15.5–101.7), Hyracoidea (33.9, 90% CI = 1.0–82.1), Cingulata (18.7, 90% CI = 10.0–27.6), Lagomorpha (12.2, 90% CI = 9.9–15.3), and Rodentia (11.9, 90% CI = 10.6–13.5; Supplementary Fig. S8b).

Beyond overall counts, predictions also revealed preferential host–virus family combinations that were not always among the most frequently observed associations. This is also the case for mammalian orders considered to be among the most important hosts of zoonotic viruses: rodents, bats, and primates¹². For example, rodents (order

Rodentia) are a fairly well-studied mammal order for almost all viral families due to their zoonotic potential¹¹, especially *Flaviviridae*, *Hantaviridae*, and *Arenaviridae*. Our predictions showed that the highest number of associations (ca. >50% rodent species affected in all cases) includes some of these families—*Flaviviridae* and *Hantaviridae*—but also some families with limited observed associations, such as *Paramyxoviridae* and *Herpesviridae* (Supplementary Fig. S9). Likewise, bats (order Chiroptera) are considered highly likely hosts of zoonotic viruses³⁹ and have well-known associations with *Paramyxoviridae*, *Coronaviridae*, and *Rhabdoviridae*, which was also confirmed by our predictions (>50% bat species affected in all cases; Supplementary Fig. S9). However, we also predicted a high percentage of bat species associated with *Flaviviridae*, *Herpesviridae*, *Astroviridae*, *Adenoviridae*, and *Picornaviridae* (all of which have a more limited number of observations compared to the former three families). Finally, primates (order Primates) were predicted to have a significantly broader viral range than currently documented, with over 50% of species expected to associate with up to 15 viral families (Supplementary Fig. S9).

On the virus side, certain viral families were consistently assigned with higher predicted probabilities, resulting in disproportionately large numbers of new host associations (Supplementary Fig. S8c). This was the case for *Flaviviridae* (median 3507 expected predicted associations, 90% CI = 2807–4173), *Paramyxoviridae* (median 3378, 90% CI = 2625–4069) and *Herpesviridae* (median 2834, 90% CI = 1950–3458). Viral families which showed the highest fold increase in the total number of associations included *Hepeviridae* (median fold-increase 12.8, 90% CI = 7.2–20.4), *Astroviridae* (median 11.5, 90% CI = 9.2–15.8), *Circoviridae* (median 11.0, 90% CI = 6.2–16.3), and *Peribunyaviridae* (median 10.9, 90% CI = 7.5–16.3; Supplementary Fig. S8d).

Key zoonotic viral families, such as *Coronaviridae* and *Paramyxoviridae*, showed shifts in association patterns when comparing observed vs predicted associations. The *Coronaviridae* family includes several viruses known to infect a wide range of animals, and notable zoonotic viruses such as SARS-CoV, MERS-CoV, and SARS-CoV-2. These viruses are known for their ability to cause respiratory, gastrointestinal, and neurological diseases in both humans and animals. We found that *Coronaviridae* had a high proportion of predicted associations across several mammal orders (Supplementary Figs. S10, S11), which included well-known hosts such as bats as well as relatively less known hosts such as ungulates (orders Perissodactyla and Cetartiodactyla). The relative difference between observed vs. predicted associations led to a substantial change in the global patterns of host species richness for *Coronaviridae* (Supplementary Fig. S12). *Paramyxoviridae* is a family of enveloped, single-stranded RNA viruses known for infecting a wide range of hosts, including humans, mammals, and birds. Notable zoonotic viruses within *Paramyxoviridae* include Nipah and Hendra, two bat-associated viruses that emerged in the Indo-Pacific region with dramatic public health consequences. Our predictions indicate that a major proportion of bat and primate species are potentially susceptible to these viruses, with median expected proportions of 0.87 for Chiroptera (90% CI = 0.80–0.95) and 0.85 for Primates (90% CI = 0.72–0.92; Supplementary Figs. S10, S11). As these are among the most speciose mammalian orders, the confirmation of these predictions could drastically alter our spatial perception of zoonotic risk (Supplementary Fig. S12).

Overall, these predictions revealed a much broader picture of mammal–virus association compared to what is currently documented (Fig. 3). Several of the established mammal–virus association patterns were confirmed, yet numerous additional ones were predicted at similar or higher levels despite limited current observational evidence. The number of known and new predicted associations was positively correlated across mammalian orders (median Pearson's $r = 0.57$, 90% CI = [0.38, 0.78]; Supplementary Fig. S13a) and viral families ($r = 0.89$, 90% CI = [0.83, 0.94]; Supplementary Fig. S14a). Yet, residual analysis

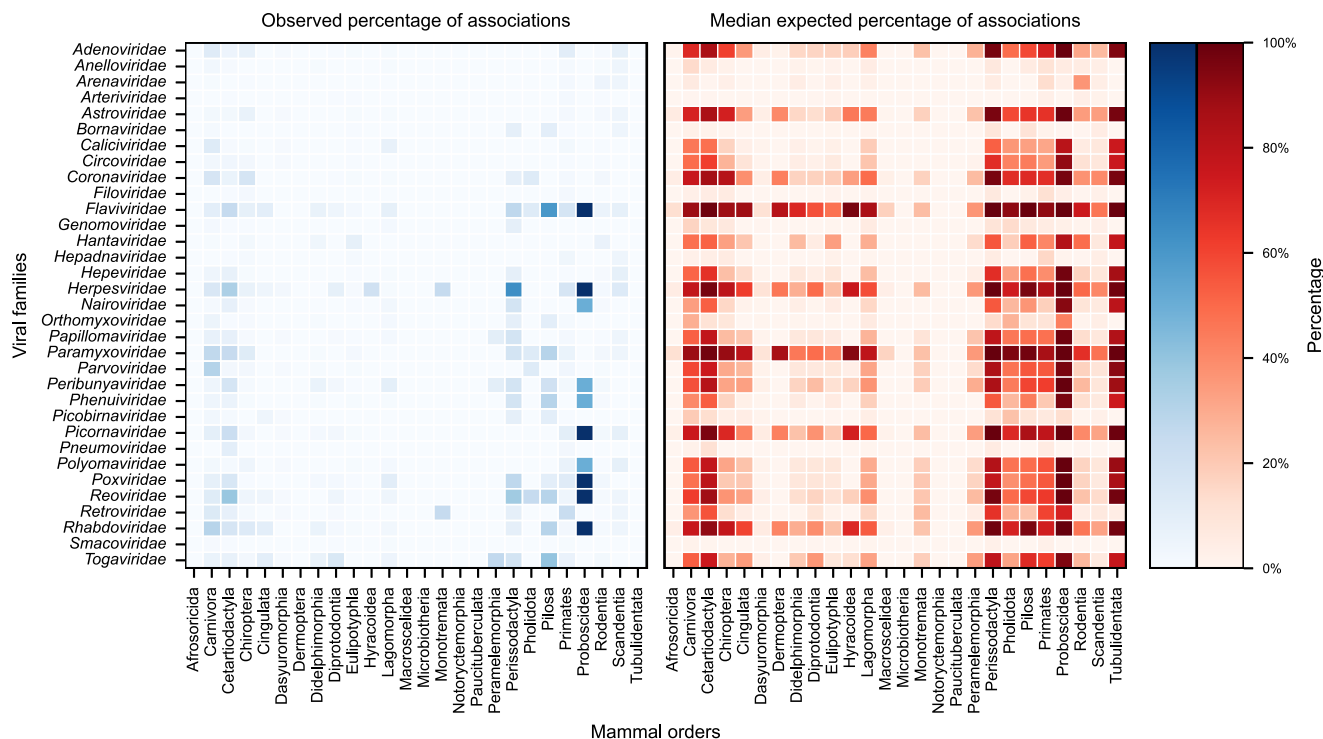


Fig. 3 | Observed vs. predicted mammal–virus associations across mammalian orders and viral families. Heatmap of observed (left) and total predicted (right) associations between mammal orders and viral families (i.e., percentage of

mammal species in the order associated with the viral family). Predicted percentages represent the median expected percentage of associations across 100 model runs.

indicated that the model captured additional information beyond the sole number of known associations, suggesting that predictions are not fully explained by the current volume of documented associations alone (Supplementary Figs. S13b, S14b). Mammalian orders with fewer known associations exhibited higher fold increases in associations (median Pearson's $r = -0.39$, 90% CI = $[-0.53, -0.30]$), indicating proportionally larger predicted knowledge gaps in poorly sampled clades (Supplementary Fig. S13c, d). Similarly, while viral families displayed roughly constant fold increases (median Pearson's $r = 0.06$, 90% CI = $[-0.11, 0.31]$), this gives credible evidence that viruses with sparser observational data were not systematically suppressed by the model (Supplementary Fig. S14c, d). Together, these results suggest that current surveillance efforts may be severely underestimating the true breadth of mammal–virus associations, particularly for poorly sampled host clades, where potentially important associations may remain undetected.

Discussion

Our DPU approach was effective at predicting missing associations from incomplete and biased observations, as demonstrated on new labels and synthetic data. In addition to identifying the most likely missing mammal–virus associations, DPU learning can also be used to quantify prediction uncertainty and research effort, expressed as propensity scores (Supplementary Fig. S15), providing a transparent basis for interpreting the results and prioritizing future efforts. The results obtained also highlight the suitability of GNNs for efficiently extracting complex emerging patterns that arise from the joint integration of mammalian phylogeography and the observed mammal–virus association networks. Importantly, the GNN allowed us to unify and extend major host–pathogen modelling traditions within a single machine learning framework. By leveraging observed association network topology^{13,24,25,40}, species traits^{11,23,26}, and evolutionary and biogeographic constraints on host–virus associations^{22,41–44}, GNNs can learn node representations that jointly encode and integrate

information from all these heterogeneous data sources, enabling the model to capture cross-modal interactions that would be difficult to recover with traditional models.

From a PU learning perspective, the principal contribution of DPU learning lies in the joint modelling of the ground truth and labelling mechanisms through the explicit conditioning of both the classifier and propensity score model on the set of currently labelled associations. Unlike traditional PU approaches^{30,33}—where predictions condition on fixed covariates, latent parameters, or predefined relational structures, and labelled associations inform model outputs solely through learned weights and parameters—DPU learning enables both components to dynamically update their predictions as new labelled associations are incorporated during the inference. This formulation confers two key advantages. First, it enables both components to capture and exploit structural dependencies across labeled associations, such as co-occurrence and shared connectivity patterns, to improve predictive performance. Second, it supports an explicitly dynamic interpretation of both the ground truth and labeling mechanisms: as new associations are discovered, predictions are reconditioned on the expanded label set, enabling iterative reinterpretation of both the underlying interaction network and the observational sampling process.

Conditioning on the currently labelled association set permits direct exploitation of the observed mammal–virus network for extracting informative patterns. Spatial or taxonomic clustering of records can inform the modeling of differential sampling, whereas higher-than-expected co-occurrence of viral families among host species can be leveraged to infer shared susceptibilities or overlapping host ranges. Notably, this approach enables the use of message-passing architectures³⁶, designed for complex graph topologies^{35,45–48}, to propagate signals through the observed association network and learn node representations that encode both its global and local structure. To integrate predefined relational structures within a unified multigraph framework, signal propagation was further extended

through phylogeographically related hosts. This facilitates generalization to species with few or no observed associations by leveraging known phylogenetic and biogeographic affinities, a key requirement for global-scale inference where the majority of mammalian species remain unstudied. For instance, a viral family with associations concentrated in a particular host clade can propagate signal simultaneously through the association network—influencing node representations of co-occurring viral families and known host species—and through the phylogenetic and biogeographic graphs to taxonomically related or geographically overlapping hosts. This allows for predictions that integrate known epidemiological context, evolutionary relatedness, and geographic proximity in a single inference step. Capturing such patterns from the observed network is critical in host–pathogen applications, where node-level features and pre-defined relational structures (e.g., mammalian and viral traits and phylogeography) alone are often insufficient to obtain a complete characterization of association probabilities²⁴.

Together, our results make DPU learning a robust and flexible foundation for applying graph-based deep learning techniques to link prediction in partially observed, biased networks. In particular, although we developed the DPU formalism for predicting missing associations in host–pathogen networks, we argue that it also applies to other network reconstruction tasks such as recommendation systems^{49,50}, drug–drug^{51,52}, drug–disease⁵³ or protein–protein interactions^{54–56}, plant–pollinator networks^{57,58}, ecological interaction networks^{59,60}, and species geographic-presence networks⁶¹. In these contexts, associations with understudied interactors, such as rare diseases, are more underrepresented in available data, not because they are unlikely to exist, but because limited study effort or experimental difficulty reduces their likelihood of being observed. DPU offers a principled way to account for such uneven gaps in the observed network, leading to more reliable and unbiased link-predictions with quantified uncertainty estimates.

Global-scale viral discovery efforts are crucial for the establishment of pandemic preparedness strategies^{3,6}. However, with hundreds of thousands of zoonotic viruses potentially unknown³, discovery programmes require economic and temporal resources which are not currently available⁶². Integrating advanced analytical tools into existing viral discovery initiatives can help identify research priorities across host taxonomies and geographic areas and anticipate potential risks from human encroachment into natural habitats^{16,63}. Machine learning tools enable the modelling of high-dimensional relationships, which makes them especially valuable when the underlying mechanisms of a system are complex and poorly understood⁶⁴, as is the case with mammal–virus associations. However, the effectiveness and efficiency of such models depend heavily on the quality of the data, the appropriateness of the learning algorithms used, and careful consideration of observational biases and data deficiencies⁶⁵.

In the context of mammal–virus associations, data are sparse, biased, and largely incomplete. Nevertheless, with rigorous training and validation, sound theoretical grounding, and synergetic use of accessible biological determinants (e.g., host and viral traits, phylogeny, and biogeography), machine learning models can still yield meaningful and informative probabilistic predictions based on documented associative patterns. Explicitly accounting for the PU nature of the data and applying tailored learning algorithms is necessary for ensuring the statistical validity and biological plausibility of model outputs. While predictions remain probabilistic and come with uncertainty (Supplementary Fig. S15), they can serve as valuable decision-support tools that may complement conventional methods for viral discovery. Results from DPU applications have practical value in this context. Our findings highlight mammalian taxa with disproportionately larger numbers of viral associations compared to currently observed levels, as is the case for Afrosoricida, where median expected viral diversity is 50 times higher than known diversity. These

taxa warrant priority focus in terms of viral discovery and local surveillance, representing potential blind spots of viral hazard^{21,24}. Validation on new labels further confirms the ability of DPU learning to predict unknown associations before these are discovered. This was the case for true insectivores (order Eulipotyphla), a known under-sampled taxon⁶⁶ which had ~76% of their newly recorded associations predicted by DPU with probabilities above 0.5.

While our results illustrate the practical value of DPU, the interpretation of predictions requires careful consideration of two methodological constraints. First, while DPU quantifies the variability across model ensembles in the expected values of aggregated predictions (e.g., the total number of associations for a specific taxon or geographic tile), these intervals do not constitute a complete measure of predictive uncertainty. In fact, estimating full predictive uncertainty would require computing joint probability distributions and accounting for correlations among single-point predictions, which is not possible in the current framework. While confidence intervals for single-point predictions (i.e., for individual mammal–virus couples) reflect complete aleatoric and epistemic predictive uncertainty, intervals for aggregated predictions only reflect how much our estimates shift under different inferred patterns of research effort (i.e., propensity scores). As such, they should be interpreted as a lower bound of true predictive uncertainty. Second, as the DPU approach naturally prioritizes unlabeled associations that resemble known positives (Supplementary Equation S67), mammal–virus couples that are highly isolated or under-sampled may receive consistently low association probabilities across all models in the ensemble. This implies that in severely underexplored regions of the mammal–virus network, the method may underestimate both the likelihood of true associations and the corresponding uncertainty. Yet, we also find that understudied taxa exhibit higher prediction uncertainty than better-studied taxa (Supplementary Figs. S16–S18). This suggests that the DPU's expected tendency to be overconfident for poorly studied taxa is, at least partially, mitigated. Furthermore, results from 10-fold cross-validation and new labels indicate that DPU achieves comparable accuracy across species regardless of how extensively their viral associations have been documented, highlighting its ability to recover true missing associations even among severely under-sampled species and those with no prior viral records (Supplementary Figs. S1, S5).

Emerging machine learning tools, such as DPU learning, can play a crucial role in optimizing resource allocation for the discovery of host–pathogen associations, supporting cost–benefit assessments, and maximizing viral discovery per effort unit. While this study focuses on viral family-level predictions, the DPU approach provides a general methodological framework that can be deployed to finer taxonomic resolutions, for example, to model host associations with viral genera of interest, provided sufficiently detailed data is available. While we considered generic viral features for predictions at the viral family level, DPU learning applications at finer levels would enable the usage of more specific features reflecting viral structure and composition. Further improvements may also involve expanding the model to account for viral associations with other key hosts (e.g., birds) as well as vectors such as mosquitoes, ticks, and sandflies⁶⁷. Incorporating these groups would allow the model to identify broader biogeographic patterns of susceptibility and provide a more comprehensive and realistic view of the global host–virus network. Additionally, as scientific knowledge on viral hosts continues to grow, predictive accuracy is expected to improve. This opens the door to adaptive learning, where model predictions guide data acquisition efforts, and newly acquired data are iteratively used to refine and update those predictions⁶⁸. In particular, further research on the stability of the predictions in response to new data can help both quantify model uncertainty and identify host–virus associations with the highest impact in reshaping the inferred network⁶⁹. Prioritizing these associations offers a

principled strategy to reduce predictive uncertainty and guide targeted surveillance in a cost-effective way.

Approaches such as DPU learning, which explicitly disentangle biological signal from sampling bias and quantify uncertainty at each stage, offer decision-makers clearer, more accountable evidence to support One Health strategies compared to standard approaches. Still, integrating explainable modeling frameworks within the DPU workflow can further strengthen this transparency by clarifying the biological and sampling-driven determinants of predicted host–virus associations⁷⁰, enabling researchers and policymakers to better evaluate model-guided prioritization. Incorporating such predictive analyses into the design of viral surveillance can transform global One Health strategies of pandemic preparedness, shifting the focus from reactive crisis response to proactive, data-driven prevention.

Methods

Input data

We retrieved known mammal–virus associations from the Global Virome in One Network (VIRION)³⁷ database, retaining associations between terrestrial mammals and viruses supported by any type of evidence (e.g., sequencing, PCR, serology). These may include viruses associated with the host’s microbiota (e.g., as hypothesized for *Smacoviridae*⁷¹). Viral associations of each mammal species were aggregated at the level of viral family to mitigate sparsity and detection bias at lower taxonomic categories, while preserving biologically meaningful patterns of host–virus compatibility¹⁴. By focusing on family-level, rather than on species-level, viral associations, we avoided the risk of having a small number of well-studied viruses dominating the dataset. The value of our inference is coarser compared to a species-level prediction, but this coarser resolution allows us to make a generic inference that also applies to poorly studied and even potentially undiscovered viruses. Viral families with fewer than 10 known associations were excluded to ensure sufficient data for robust model training and reliable prediction performance, and because they lacked enough data for meaningful validation. Also, we removed *Homo sapiens* and all mammal species for which mammalian traits, phylogenetic information, and geographic distribution (see below) were not available. This filtering produced a dataset of 5330 mammal species $m \in M$ (out of 5642 terrestrial mammal species officially recognized by IUCN (2020)⁷²) and 33 viral families $v \in V$ linked by 4968 unique associations; 686 additional associations from GenBank³⁸ were incorporated as new labels. Here, M and V denote the sets of mammal species and viral families, respectively. Mammal species names were harmonized across different data sources following the IUCN (2020) binomial nomenclature.

New mammal–virus associations (new labels) were systematically collected from GenBank using a custom pipeline in R (packages: *rentrez*⁷³, *rvest*⁷⁴, *xml2*⁷⁵). The search targeted viral records for all mammalian species, following IUCN 2020 taxonomy, deposited between January 2022 and September 2024 in the Nucleotide database (<https://www.ncbi.nlm.nih.gov/genbank/>). Associations were based on PCR amplification from host samples. We filtered for novel viral family–mammal associations not previously present in our training data to ensure these new labels provided an independent validation set.

Mammal and viral features. Each mammal species $m \in M$ was associated with a vector x_m of 18 mammal traits sourced from the COMBINE and PHYLACINE v.1.2 datasets^{76,77}. These encompass species’ adult body mass (log-transformed), maximum longevity, gestation length, interbirth interval, weaning age (log-transformed), mean litter size, annual litter output, trophic level, and the first 10 phylogenetic eigenvectors. These traits are known to capture biological mechanisms underlying host susceptibility and pathogen transmission^{14,44,78–80}. Each viral family $v \in V$ was associated with a vector x_v of 7 binary viral

features sourced from Wardeh et al. 2021²⁴, encompassing viruses’ nucleic acid type (DNA/RNA), retro-transcription capability, strand structure, genome topology (circular/linear), genome segmentation, and envelope presence. These viral traits are extensively used in virology for virus classification and characterization⁸¹ and have already been proven to be important for predicting viral host-range^{24,82–90}. A detailed description of the mammalian and viral traits used in the analysis, together with the rationale for their selection, is provided in the Supplementary Information file, Supplementary Subsections A.1.1 and A.1.2.

Mammals reciprocal relationships. Mammalian phylogenetic and biogeographical relationships were also considered to inform our model. Phylogenetic relatedness was modeled using patristic distances (extracted from the PHYLACINE v. 1.2 database⁷⁷), that is, the elapsed time since the most recent common ancestor for a pair of species^{91,92}. Biogeographic relationships were estimated by considering the spatial overlap between any two pairs of co-occurring species (based on IUCN Red List v. 2020-2 range maps⁷²). Pairwise range overlaps were computed using the “sf” R package⁹³. While phylogeny captures the evolutionary relationships of mammal species, biogeography reflects their geospatial connectivity⁹⁴. Both represent two of the most important determinants of interspecific viral transmission, quantitatively linked to pathogen sharing within animal communities^{24,41–44,95}.

Modelling approach

The task of predicting missing associations between mammal species and viral families was framed as a dynamic positive-unlabeled (DPU) problem. Documented mammal–virus associations were treated as labeled cases; all others were considered unlabeled.

Graph representation of mammal–virus and mammal–mammal interactions. Current knowledge of mammal–virus associations was represented using a bipartite and unweighted graph referred to as the observed network of mammal–virus associations $G_{MV} = (M \cup V, E_{MV})$. The node set $M \cup V$ consisted of all mammalian species and viral families included in the dataset. The edge set $E_{MV} \subset M \times V$ comprised all mammal–virus pairs with a recorded association. In Supplementary Fig. S19, we reported the degree distribution for mammal species and viral families in G_{MV} . While both the underlying biology and the sampling process contribute to this distribution, substantial evidence indicates that the observed heterogeneity is predominantly driven by differences in research effort^{12,18,40}, highlighting how some species and viral families have been far more extensively studied than others. This network representation offered a compact and structured way to encode all known mammal–virus associations (representing our primary source of information about mammalian viral diversity and virus–host-range) within a graph-theoretic framework. This allowed GNNs to directly leverage the network structure underlying known associations in order to characterize mammal species and viral families based on their observed associations and to uncover global patterns and local motifs to improve the plausibility of missing links by using contextual information derived from currently labeled data.

Additionally, we defined a mammalian phylogenetic graph to represent species phylogenetic distances among species, and a mammalian biogeographical graph to represent their geographical adjacencies and cooccurrences (Supplementary Information file, Supplementary Subsection A.2.1). Modeling phylogenetic and biogeographical relationships among mammals is crucial for uncovering latent patterns, biological constraints and dependencies in our data. This approach provides deeper insight into phylogenetically and biogeographically conserved patterns of host association within viral families and helps compensate for observational data gaps by

leveraging known interspecific relationships to propagate viral information through hosts.

Overcoming observational biases. Four main sources of observation biases were identified and explicitly modeled in our DPU framework (Supplementary Information file, Supplementary Subsection A.2.2). These biases include: human-centric reporting⁹⁶, arising because knowledge of infectious agents is concentrated on humans and domesticated animals rather than wild species^{12,97}; geographic biases^{17,98}, due to higher research output in wealthier countries, leading to better knowledge of associations relevant to those regions; biological biases⁹⁹, whereby virus–host associations with more visible effects are more likely to be studied; and self-reinforcing biases¹⁸, in which hosts already known to be susceptible to certain pathogens are more likely to be studied when a new agent of that type is discovered (e.g., coronaviruses in bats).

These patterns are captured in our model through data-driven proxies for research effort that integrate network-derived, geographic, and taxonomic signals, capturing how, for each viral family, research effort accumulates and propagates across related hosts. The number of known viral associations per host serves as a proxy for overall research intensity, subsuming self-reinforcing biases whereby hosts already known to harbor certain pathogens usually attract disproportionate subsequent studies, as well as human-centric biases concentrating knowledge on economically or medically relevant species. Host taxonomic structure captures both biological biases—wherein phylogenetically related hosts share susceptibility to the same viral families—and self-reinforcing biases that redirect sampling effort toward relatives of known reservoir species. Mean host exposure to field studies reflects geographic and human-centric sampling effort, as hosts occurring in regions with higher research output or in proximity to human populations are more likely to be included in viral surveys. Finally, viral taxonomy accounts for research history and biological biases, capturing the uneven attention directed toward viral families with greater perceived zoonotic potential or more conspicuous pathological effects.

DPU formalism. The DPU framework generalizes the SCAR³⁰ and SAR³³ assumptions by introducing a dynamic labeling mechanism, in which past observations inform both current predictions and propensity score estimation. This requires two complementary models: a classifier (Supplementary Equation S9) that predicts the likelihood that a given association exists and a propensity score model (Supplementary Equation S11) that estimates the likelihood of an existing association getting labeled—both conditioned on the labels of all other associations. Importantly, to prevent data leakage, predictions for each focal association are generated by masking its own label from both components, ensuring that neither model has access to it, but only to the labels of all other remaining associations.

The labeled or unlabeled status of a potential association is nonetheless informative during inference (Supplementary Equation S12). By the PU learning convention, a labeled association is always treated as a confirmed positive. Conversely, an association that remains unobserved despite a high propensity score (i.e., a high probability of detection had it existed) is less likely to represent a true link, even if the classifier deems it biologically plausible. In contrast, a high classifier output combined with a low propensity score is indicative of a likely missing association. The DPU framework leverages both components to make principled probabilistic inferences about unlabeled associations: classifier outputs discriminate biologically plausible positives from negatives, as in conventional models, while propensity scores provide complementary information by distinguishing associations with low research effort—representing genuinely unknown cases—from those with high research effort—which constitute more reliable negatives. Jointly integrating these two

components allows the DPU framework to account for both biological plausibility and observational evidence during inference. A full treatment of the DPU formalism is provided in Supplementary Methods, Subsection A.2.3.

Bayesian DPU modelling. For every viral family $v \in V$, our Bayesian DPU model consisted of two hierarchical components—one functioning as a classifier (with parameters Θ_v) and the other as a propensity score model (with parameters Φ_v). To ensure consistency, all classifiers shared a common architecture and parameters' prior distributions across viral families, and the same applied to the propensity score models. Parameters Θ_v and Φ_v were structured hierarchically following mammalian taxonomy, estimating parameters at multiple taxonomic levels and shared within related taxonomic groups. This structure allowed us to borrow statistical strength from data-rich groups to improve estimates in data-poor groups, by generalizing patterns observed in well-studied clades to make inferences in related but less-studied ones.

Given the database of associations' labels K (see Supplementary Equation S7), we computed the probability distribution of the models' parameters:

$$Pr(\Theta, \Phi|K) = \prod_{v \in V} Pr(\Theta_v, \Phi_v) \prod_{m \in M} p_{(m,v)}^{S(m,v)} (1 - p_{(m,v)})^{1-S(m,v)}, \quad (1)$$

where we assumed prior independence of parameters Θ_v, Φ_v associated with different viral families, and that the binary labels $S(m, v) \in \{0, 1\}$ in K are generated by a Bernoulli process with success probability

$$\begin{aligned} p_{(m,v)} &= Pr(S(m, v) = 1 | K_{M \times \{v\}}, \Theta_v, \Phi_v) \\ &= Pr(Y(m, v) = 1 | K_{M \times \{v\}}, \Theta_v) \\ &\quad \cdot Pr(S(m, v) = 1 | K_{M \times \{v\}}, Y(m, v) = 1, \Phi_v) \end{aligned} \quad (2)$$

Here, $Pr(\Theta_v, \Phi_v)$ is the prior of the models' parameters and $\prod_{m \in M} p_{(m,v)}^{S(m,v)} (1 - p_{(m,v)})^{1-S(m,v)}$ is the likelihood for data related to viral family v and $K_{M \times \{v\}}$ is the database of associations' labels where all the labels of all potential associations with v have been masked. Finally, $Y(m, v) \in \{0, 1\}$ is the true value of couple (m, v) , i.e., $Y(m, v) = 1$ if (m, v) is a true positive and $Y(m, v) = 0$ if (m, v) is a true negative. The classifier was defined as

$$Pr(Y(m, v) = 1 | K_{M \times \{v\}}, \Theta_v) = \pi_{f(m), v}, \quad (3)$$

with $\pi_{f,v}$ being a parameter within Θ_v . We modeled propensity scores as

$$\begin{aligned} Pr(S(m, v) = 1, | K_{M \times \{v\}}, Y(m, v) = 1, \Phi_v) &= \\ &= \frac{1 + E_v \cdot g_{m,v}(K_{M \times \{v\}})}{1 + E_v \cdot g_{m,v}(K_{M \times \{v\}}) + e_{m,v}(K_{M \times \{v\}}, \Phi_v)} e_{m,v}(K_{M \times \{v\}}, \Phi_v), \end{aligned} \quad (4)$$

with

$$\begin{aligned} e_{m,v}(K_{M \times \{v\}}, \Phi_v) &= \\ &= C_{f(m), v} + \frac{\sigma\left(\frac{d_{m,v}(K_{M \times \{v\}}) - A_{f(m), v}}{B_{f(m), v}}\right) - \sigma\left(-\frac{A_{f(m), v}}{B_{f(m), v}}\right)}{1 - \sigma\left(-\frac{A_{f(m), v}}{B_{f(m), v}}\right)} D_{f(m), v} (1 - C_{f(m), v}) \end{aligned} \quad (5)$$

where $A_{f,v}, B_{f,v}, C_{f,v}, D_{f,v}, E_v$ are parameters in Φ_v , $f(m)$ is the mammalian family of m , $g_{m,v}(K_{M \times \{v\}})$ represents a proxy for geographic sampling bias defined as

$$g_{m,v}(K_{M \times \{v\}}) = \frac{1}{|V| - 1} \sum_{v' \in V \setminus \{v\}} \left(\max_{m' \in M \setminus \{m\}} g(m', v') \right), \quad (6)$$

with:

$$g(m, m', v') := \begin{cases} \frac{\text{overlap}(m, m')}{\text{overlap}(m', m')} & \text{if } S(m', v') = 1, \\ 0 & \text{otherwise,} \end{cases} \quad (7)$$

and $d_{m,v}(K_{M \times \{v\}})$ is the degree of mammal species m in the masked database $K_{(m,v)}$:

$$d_{m,v}(K_{M \times \{v\}}) = \begin{cases} d_m(K) - 1 & \text{if } S(m, v) = 1, \\ d_m(K) & \text{otherwise.} \end{cases} \quad (8)$$

All the parameters defined at a mammalian family-level f were drawn from a distribution parametrized by the corresponding parameters at the mammalian order-level $o(f)$. In turn, the parameters defined at a mammalian order-level o were drawn from a distribution parametrized by parameters at the global mammalian level. More information about the Bayesian model architecture, the functional dependences between the models' parameters, the model's priors, and the overall model's rationale can be found in the Supplementary Information file, Supplementary Subsection A.2.4.

Calculations were performed using the TeraStat2 HPC¹⁰⁰ using the JAGS¹⁰¹ library interfaced with Python¹⁰² through the package PyJAGS¹⁰³. Sampling of the model's parameters 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 tests¹⁰⁴, specifically when R-hat values were <1.01 and ESS values were >1000 for all parameters and hyperparameters of the model. Using this strategy, the resulting chains contained between 7000 and 32,000 samples for each viral family, depending on convergence time, which ranged, using TeraStat2 HPC, from 3 to 18 h across families.

GNN classifier model. Two distinct mammal–virus latent adjacency matrices were constructed to capture the positional context of each node within the multi-graph structure of mammal–virus interactions: one based on mammalian phylogenetic relationships, and one on their biogeographic distribution. Weighted edges were assigned between a mammal species $m \in M$ and viral family $v \in V$ by aggregating observed associations with v across mammal species either phylogenetically or biogeographically related to m . In practice, the edge weight is computed as a normalized weighted sum: the number of observed associations with viral family v among related mammal species is combined with the strength of their relationships to m through a fixed (i.e., non-learnable) mammal–mammal interaction matrix. Formally, given a normalized mammal–virus adjacency matrix $A_{MV}^{(G)}$ induced by observed associations in a given mammal–virus network G ,

$$A_{MV}^{(G)}[m, v] = \begin{cases} \frac{1}{\sqrt{d_m^{(G)} d_v^{(G)}}} & \text{if } (m, v) \in E(G), \\ 0 & \text{otherwise,} \end{cases} \quad (9)$$

and a predefined mammal–mammal interaction network $M: M \times M \rightarrow [0, 1]$, we define the latent mammal–virus adjacency matrix constructed via G and M as

$$Q_{MV}^{(M, G)} = \left(D_{MA_{MV}^{(G)}}\right)^{-1} M A_{MV}^{(G)}. \quad (10)$$

Unlike adjacency matrices derived directly from empirical data, these matrices incorporate meta-edges that connect mammal species to viral families even in the absence of documented associations. Specifically, a mammal species can be indirectly linked to a viral family via phylogenetically or biogeographically related species that are directly linked to it (i.e., known hosts). These meta-paths, inspired by the method used in Graph Transformer Networks⁴⁸ for node representation learning on heterogeneous graphs, traverse both the observed mammal–virus and the fixed mammal–mammal networks,

allowing the model to capture latent structural dependencies across the two. By constructing such latent matrices, we can use message-passing using the latent adjacency matrix to generate node embeddings that reflect both mammal–virus known associations and meta-interactions, enabling inference in mammalian groups where direct host–virus associations are sparse or absent by leveraging known interspecific relationships.

Message-passing was performed using MLP with one hidden layer (see Supplementary Equations S53–S56). The resulting node embeddings from the two matrices were aggregated through a weighted SoftMax Aggregation¹⁰⁵ for each mammal and virus node (Supplementary Equation S58), providing a comprehensive representation of the nodes that reflects both phylogenetic and biogeographic relationships among mammals, observed (i.e., labeled) mammal–virus associations, and node-specific mammalian and viral features. Finally, these embeddings were concatenated for each mammal–virus and processed through a Dense Neural Network to compute the probability of existent association between them. Non-linearity in the model's layers was introduced via PReLU activation functions¹⁰⁶, and each GNN layer was normalized using GraphNorm¹⁰⁷. The model was implemented in Python¹⁰², using the Pytorch¹⁰⁸ and Pytorch Geometric¹⁰⁹ libraries. All training and inference were performed on NVIDIA V100 GPUs with 32 GB of memory, with each model requiring ~17 h to train. More detailed information about the GNN architecture can be found in the Supplementary Information file, Supplementary Subsection A.2.5.

Model training, validation, and predictions

In this section, we outline the methodological pipeline used to train our models, estimate their predictive performance, and generate final predictions with quantified uncertainty.

Training procedure. First, we used the Bayesian DPU model to estimate reasonable joint probability distributions of the propensity scores across all potential mammal–virus associations in our dataset. Next, we either sampled configurations of propensity scores or used their expected values to define a non-negative DPU empirical risk R_ℓ^{DPU} . This served as the loss function to train the GNN classifier via gradient descent with the AdamW optimizer^{110,111}. In particular, given a graph-dataset $D = \left\{ \left(G^{(k)}, \mathcal{E}_S^{(k)}, \mathcal{E}_U^{(k)} \right) \right\}_{k=0}^K$, with $G^{(k)}$ being a subgraph of the observed mammal–virus association network, $\mathcal{E}_S^{(k)}$ being the set of labeled associations missing from $G^{(k)}$, and $\mathcal{E}_U^{(k)}$ being a set of unlabeled associations, we defined the DPU risk R_ℓ^{DPU} conditioned to a set of propensity scores $\mathbf{e} = \{e_{(m,v)}\}_{(m,v) \in M \times V}$ as:

$$R_\ell^{\text{DPU}}[\mathbf{g}, \mathbf{e}, D] := \frac{1}{|V|} \sum_{v \in V} R_{\ell,v}^{+,S}[\mathbf{g}, \mathbf{e}, D] + \max \left\{ 0, R_{\ell,v}^{-,S}[\mathbf{g}, \mathbf{e}, D] + R_{\ell,v}^{-,U}[\mathbf{g}, D] \right\}, \quad (11)$$

with:

$$R_{\ell,v}^{+,S}[\mathbf{g}, \mathbf{e}, D] = \frac{\pi_{s,v}}{\sum_{k=1}^K |\mathcal{E}_{S,v}^{(k)}|} \sum_{k=1}^K \sum_{x \in \mathcal{E}_{S,v}^{(k)}} \frac{1}{e_x} \ell(\mathbf{g}(x, G^{(k)}), 1) \quad (12)$$

$$R_{\ell,v}^{+,S}[\mathbf{g}, \mathbf{e}, D] = \frac{\pi_{s,v}}{\sum_{k=1}^K |\mathcal{E}_{S,v}^{(k)}|} \sum_{k=1}^K \sum_{x \in \mathcal{E}_{S,v}^{(k)}} \left(1 - \frac{1}{e_x} \right) \ell(\mathbf{g}(x, G^{(k)}), 0), \quad (13)$$

$$R_{\ell,v}^{-,U}[\mathbf{g}, D] = \frac{1 - \pi_{s,v}}{\sum_{k=1}^K |\mathcal{E}_{U,v}^{(k)}|} \sum_{k=1}^K \sum_{x \in \mathcal{E}_{U,v}^{(k)}} \ell(\mathbf{g}(x, G^{(k)}), 0). \quad (14)$$

Here: $\mathcal{E}_{S,v}^{(k)} = \mathcal{E}_S^{(k)} \cap \{(m,v) | m \in M\}$, $\mathcal{E}_{U,v}^{(k)} = \mathcal{E}_U^{(k)} \cap \{(m,v) | m \in M\}$, and $\pi_{s,v}$ is the proportion of labeled hosts with v . Finally, $\ell(\hat{y}, y) = -y \log(\hat{y}) - (1 - y) \log(1 - \hat{y})$ is the binary cross entropy. The proposed loss function is built on recent advancements in developing learning algorithms for training bias-free predictive models in PU scenarios^{33,112}. Minimizing DPU risk encourages the model to treat labeled associations as positive instances without strictly enforcing that all unobserved associations are true negatives, using propensity scores to debias predictions (see Supplementary Information file, Supplementary Subsection A.3.1).

Mixed precision-training was used to improve computational efficiency¹¹³ and propensity scores were clipped to 0.005 during the training for limiting unstable outcomes¹¹⁴. Finally, predictions for unlabeled associations were obtained by combining their respective propensity scores with the GNN classifier's outputs (Supplementary Equation S11).

Cross-validation, external evaluation, and synthetic benchmarking.

We evaluated the performance of the DPU approach using three complementary strategies. First, we performed a cross-validation by dividing all labeled and unlabeled associations (from the VIRION database) into 10 non-overlapping balanced folds (Supplementary Information file, Supplementary Subsections A.3.2, A.3.3), each containing ~497 labeled and 17,092 unlabeled associations. Each fold was further divided into 50 smaller subfolds, each containing ~10 labeled and 342 unlabeled associations. For each subfold, we constructed a subgraph by removing from the original mammal–virus graph only the edges corresponding to the 10 labeled associations in that subfold. This subgraph was then used to assess the model's predictions for both the held-out labeled associations and the unlabeled associations in the subfold, conditional on the remaining observed network structure. This subgraph/subfold creation process was repeated five times for each fold, resulting in 250 distinct subgraphs/subfolds per fold. During the cross-validation, each fold served eight times as part of the training set, once as the validation set for model tuning, and once as the test set for model validation. Because labeled and unlabeled associations that are used as prediction targets during the training are not contained in the validation and test folds, model performance is evaluated only on associations and network configurations that were not seen during training, ensuring an unbiased assessment across both mammal–virus pairs and observed network structures.

Second, we retrained the classifier model using all 10 folds and the hyperparameters optimized during cross-validation. Thus, we validated the model on the 686 new labels—confirmed associations absent from the VIRION database and therefore unlabeled during training—obtained from the GenBank database for external evaluation. Third, we simulated associations between mammalian species and two synthetic viral families, for which we controlled the ground truth host-range and the labelling mechanism (Supplementary Information file, Supplementary Subsection A.1.4). To define the ground truth for both synthetic viral families, we focussed on phylogenetic and spatial patterns for determining the drive of susceptibility and exposure of mammalian species to viral pathogens, as these represent the two critical factors influencing viral host range^{24,41,44,95}. By assessing the ability of the DPU framework to retrieve these synthetic associations given a subset of labelled synthetic associations, we tested whether the phylogeny-explicit and spatial-explicit components of our modelling approach were able to effectively capture the generated patterns. All the models were trained using, for each potential mammal–virus association, the expected value of its propensity score. Expected values of propensity scores were also used during inference to compute probabilities for missing associations (via Supplementary Equation S11).

Each validation strategy employed a different set of performance metrics. For nested cross-validation and new label validation—where only positive and unlabeled instances were available—we used naïve

Recall, Negative Recall, ROC-AUC, PR-AUC, Precision, and TSS (TSS = Recall + Negative Recall–1¹¹⁵). The naïve prefix is used here for all metrics to indicate that they all treat naïvely unlabeled instances as true negatives, despite the inherent uncertainty in their labeling. A novel DPU metric, SAR-PUF1 (see Supplementary Information file, Supplementary Subsection B.1), was used to assess the classifier's ability to match the ground truth¹¹⁶ using only PU data and estimated propensity scores. SAR-PUF1 was also employed for early stopping during all model training. In contrast, for synthetic datasets, where both synthetic positives and negatives were available, we evaluated prediction performance on synthetic associations unlabeled during training using standard (i.e., non-naïve) metrics: Recall, Negative Recall, ROC-AUC, PR-AUC, Precision, and TSS.

Generating final predictions and quantifying uncertainty. Finally, we used the complete observed network of mammal–virus associations (obtained by combining labels from VIRION and GenBank databases) to train both the Bayesian DPU model and an ensemble of 100 GNN classifiers, in order to generate the most up-to-date predictions. Each GNN in the ensemble was trained using a unique DPU loss function, determined by a distinct set of propensity scores sampled from the posterior distribution of the Bayesian DPU model. These same propensity scores were also used during inference to compute probabilities for missing associations (via Supplementary Equation S11). Additionally, each GNN was initialized with different random weights and trained on a distinct set of subgraphs randomly sampled from the observed network of mammal–virus associations. This ensemble approach enabled us to propagate and thereby quantify the uncertainty in the propensity scores captured by the Bayesian DPU model (and thus in our estimates of research effort distribution) into the final predictions.

For each model, we computed the expected number of associations at different aggregation levels (e.g., per host species/order, or viral family) and then used the median of these expected values across the ensemble as our primary summary statistics (i.e., median expected values). To reflect variability across ensemble members, we also reported 90% empirical ensemble intervals, calculated as the 5th and 95th percentiles of expected values across the 100 ensemble members, which defined our confidence intervals (CIs) used throughout the main text. These CIs do not represent full predictive uncertainty, as they do not capture aleatoric uncertainty or account for correlations between individual predictions in aggregated predictions (see Supplementary Information file, Supplementary Subsection B.2).

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The data generated in this study have been deposited in the Zenodo database under accession code <https://doi.org/10.5281/zenodo.20764782>¹¹⁷.

Code availability

The codes generated in this study have been deposited in the Zenodo database under accession code <https://doi.org/10.5281/zenodo.20764782>¹¹⁷.

References

1. Jones, K. E. et al. Global trends in emerging infectious diseases. *Nature* **451**, 990–993 (2008).
2. White, R. J. & Razgour, O. Emerging zoonotic diseases originating in mammals: a systematic review of effects of anthropogenic land-use change. *Mammal. Rev.* **50**, 336–352 (2020).

3. Carroll, D. et al. The Global Virome Project. *Science* **359**, 872–874 (2018).
4. Carlson, C. J., Zipfel, C. M., Garnier, R. & Bansal, S. Global estimates of mammalian viral diversity accounting for host sharing. *Nat. Ecol. Evol.* **3**, 1070–1075 (2019).
5. Mollentze, N., Babayan, S. A. & Streicker, D. G. Identifying and prioritizing potential human-infecting viruses from their genome sequences. *PLoS Biol.* **19**, e3001390 (2021).
6. Bernstein, A. S. et al. The costs and benefits of primary prevention of zoonotic pandemics. *Sci. Adv.* **8**, eabl4183 (2022).
7. Anthony, S. J. et al. A strategy to estimate unknown viral diversity in mammals. *mBio* **4**, e00598–13 (2013).
8. Wille, M., Geoghegan, J. L. & Holmes, E. C. How accurately can we assess zoonotic risk? *PLoS Biol.* **19**, e3001135 (2021).
9. Caceres-Escobar, H. et al. Operationalizing One Health: environmental solutions for pandemic prevention. *EcoHealth* **20**, 156–164 (2023).
10. Carlson, C. J. et al. The future of zoonotic risk prediction. *Philos. Trans. R. Soc. B Biol. Sci.* **376**, 20200358 (2021).
11. Han, B. A., Schmidt, J. P., Bowden, S. E. & Drake, J. M. Rodent reservoirs of future zoonotic diseases. *Proc. Natl. Acad. Sci. USA* **112**, 7039–7044 (2015).
12. Olival, K. J. et al. Host and viral traits predict zoonotic spillover from mammals. *Nature* **546**, 646–650 (2017).
13. Albery, G. F. et al. The science of the host–virus network. *Nat. Microbiol.* **6**, 1483–1492 (2021).
14. Tonelli, A., Caceres-Escobar, H., Blagrove, M. S. C., Wardeh, M. & Di Marco, M. Identifying life-history patterns along the fast-slow continuum of mammalian viral carriers. *R. Soc. Open Sci.* **11**, 231512 (2024).
15. Robles-Fernández, Á.L., Santiago-Alarcon, D. & Lira-Noriega, A. Wildlife susceptibility to infectious diseases at global scales. *Proc. Natl. Acad. Sci. USA* **119**, e2122851119 (2022).
16. Di Marco, M., Marcolin, L., Tonelli, A., Skinner, E. & Catucci, E. Macroecological approaches for the prediction of zoonotic disease risk. *Nat. Sustain.* **9**, 35–45 (2026).
17. Becker, D. J., Crowley, D. E., Washburne, A. D. & Plowright, R. K. Temporal and spatial limitations in global surveillance for bat filoviruses and henipaviruses. *Biol. Lett.* **15**, 20190423 (2019).
18. Gibb, R. et al. Mammal virus diversity estimates are unstable due to accelerating discovery effort. *Biol. Lett.* **18**, 20210427 (2022).
19. Bekker, J. & Davis, J. Learning from positive and unlabeled data: a survey. *Mach. Learn.* **109**, 719–760 (2020).
20. Iuchi, H. et al. Bioinformatics approaches for unveiling virus–host interactions. *Comput. Struct. Biotechnol. J.* **21**, 1774–1784 (2023).
21. Tonelli, A., Blagrove, M. S. C., Wardeh, M. & Di Marco, M. A framework to predict zoonotic hosts under data uncertainty: a case study on betacoronaviruses. *Methods Ecol. Evol.* **16**, 611–624 (2025).
22. Farrell, M. J., Elmasri, M., Stephens, D. A. & Davies, T. J. Predicting missing links in global host–parasite networks. *J. Anim. Ecol.* **91**, 715–726 (2022).
23. Pandit, P. S. et al. Predicting wildlife reservoirs and global vulnerability to zoonotic Flaviviruses. *Nat. Commun.* **9**, 5425 (2018).
24. 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. *Nat. Commun.* **12**, 3954 (2021).
25. Becker, D. J. et al. Optimising predictive models to prioritise viral discovery in zoonotic reservoirs. *Lancet Microbe* **3**, e625–e637 (2022).
26. Han, B. A. et al. Undiscovered bat hosts of filoviruses. *PLoS Negl. Trop. Dis.* **10**, e0004815 (2016).
27. Reiter, R. On closed world data bases. In *Logic and Data Bases* (eds Gallaire, H. & Minker, J.) 55–76 (Springer US, 1978).
28. Yang, Y., Liang, K. J. & Carin, L. Object detection as a positive-unlabeled problem. Preprint at <https://doi.org/10.48550/arXiv.2002.04672> (2020).
29. Elmasri, M., Farrell, M. J., Davies, T. J. & Stephens, D. A. A hierarchical Bayesian model for predicting ecological interactions using scaled evolutionary relationships. *Ann. Appl. Stat.* **14**, (2020).
30. Elkan, C. & Noto, K. Learning classifiers from only positive and unlabeled data. In *Proc. of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 213–220 (ACM, 2008).
31. Cressie, N., Calder, C. A., Clark, J. S., Hoef, J. M. V. & Wikle, C. K. Accounting for uncertainty in ecological analysis: the strengths and limitations of hierarchical statistical modeling. *Ecol. Appl.* **19**, 553–570 (2009).
32. Fiske, I. & Chandler, R. unmarked: an R package for fitting hierarchical models of wildlife occurrence and abundance. *J. Stat. Softw.* **43**, (2011).
33. 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* Vol. 11907 (eds Brefeld, U. et al.) 71–85 (Springer International Publishing, 2020).
34. Hamilton, W. L., Ying, R. & Leskovec, J. Representation learning on graphs: methods and applications. Preprint at <https://doi.org/10.48550/arXiv.1709.05584> (2018).
35. Dwivedi, V. P. & Bresson, X. A generalization of transformer networks to graphs. Preprint at <https://doi.org/10.48550/ARXIV.2012.09699> (2020).
36. Bronstein, M. M., Bruna, J., LeCun, Y., Szlam, A. & Vandergheynst, P. Geometric deep learning: going beyond Euclidean data. *IEEE Signal Process. Mag.* **34**, 18–42 (2017).
37. Carlson, C. J. et al. The Global Virome in One Network (VIRION): an atlas of vertebrate–virus associations. *mBio* **13**, e02985–21 (2022).
38. Clark, K., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J. & Sayers, E. W. GenBank. *Nucleic Acids Res.* **44**, D67–D72 (2016).
39. Luis, A. D. et al. A comparison of bats and rodents as reservoirs of zoonotic viruses: are bats special? *Proc. R. Soc. B Biol. Sci.* **280**, 20122753 (2013).
40. Poisot, T. et al. Network embedding unveils the hidden interactions in the mammalian virome. *Patterns* **4**, 100738 (2023).
41. Davies, T. J. & Pedersen, A. B. Phylogeny and geography predict pathogen community similarity in wild primates and humans. *Proc. Biol. Sci.* **275**, 1695–1701 (2008).
42. 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. *J. Anim. Ecol.* **83**, 671–680 (2014).
43. Stephens, P. R. et al. Parasite sharing in wild ungulates and their predators: effects of phylogeny, range overlap, and trophic links. *J. Anim. Ecol.* **88**, 1017–1028 (2019).
44. Albery, G. F., Eskew, E. A., Ross, N. & Olival, K. J. Predicting the global mammalian viral sharing network using phylogeography. *Nat. Commun.* **11**, 2260 (2020).
45. Kipf, T. N. & Welling, M. Semi-supervised classification with graph convolutional networks. Preprint at <https://doi.org/10.48550/ARXIV.1609.02907> (2016).
46. Hamilton, W. L., Ying, R. & Leskovec, J. Inductive representation learning on large graphs. Preprint at <https://doi.org/10.48550/ARXIV.1706.02216> (2017).
47. Veličković, P. et al. Graph attention networks. Preprint at <https://doi.org/10.48550/ARXIV.1710.10903> (2017).
48. Yun, S., Jeong, M., Kim, R., Kang, J. & Kim, H. J. Graph transformer networks. Preprint at <https://doi.org/10.48550/ARXIV.1911.06455> (2019).

49. Ren, Y., Ji, D. & Zhang, H. Positive unlabeled learning for deceptive reviews detection. In *Proc. 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* 488–498 (Association for Computational Linguistics, 2014).
50. Li, H., Chen, Z., Liu, B., Wei, X. & Shao, J. Spotting fake reviews via collective positive-unlabeled learning. In *2014 IEEE International Conference on Data Mining* 899–904 (IEEE, 2014).
51. Hameed, P. N., Verspoor, K., Kusljic, S. & Halgamuge, S. Positive-unlabeled learning for inferring drug interactions based on heterogeneous attributes. *BMC Bioinform.* **18**, 140 (2017).
52. Kpanou, R., Osseni, M. A., Tossou, P., Laviolette, F. & Corbeil, J. On the robustness of generalization of drug–drug interaction models. *BMC Bioinform.* **22**, 477 (2021).
53. Claesen, M., De Smet, F., Suykens, J. A. K. & De Moor, B. A robust ensemble approach to learn from positive and unlabeled data using SVM base models. *Neurocomputing* **160**, 73–84 (2015).
54. Kiliç, C. & Tan, M. Positive unlabeled learning for deriving protein interaction networks. *Netw. Model. Anal. Health Inform. Bioinform.* **1**, 87–102 (2012).
55. Song, H., Bremer, B. J., Hinds, E. C., Raskutti, G. & Romero, P. A. Inferring protein sequence-function relationships with large-scale positive-unlabeled learning. *Cell Syst.* **12**, 92–101.e8 (2021).
56. Chen, H. et al. Systematic evaluation of machine learning methods for identifying human–pathogen protein–protein interactions. *Brief. Bioinform.* **22**, bbaa068 (2021).
57. Young, J.-G., Valdovinos, F. S. & Newman, M. E. J. Reconstruction of plant–pollinator networks from observational data. *Nat. Commun.* **12**, 3911 (2021).
58. Bartomeus, I. Understanding linkage rules in plant–pollinator networks by using hierarchical models that incorporate pollinator detectability and plant traits. *PLoS ONE* **8**, e69200 (2013).
59. De Aguiar, M. A. M. et al. Revealing biases in the sampling of ecological interaction networks. *PeerJ* **7**, e7566 (2019).
60. Weinstein, B. G. & Graham, C. H. Persistent bill and corolla matching despite shifting temporal resources in tropical hummingbird–plant interactions. *Ecol. Lett.* **20**, 326–335 (2017).
61. Ward, G., Hastie, T., Barry, S., Elith, J. & Leathwick, J. R. Presence-only data and the EM algorithm. *Biometrics* **65**, 554–563 (2009).
62. Dobson, A. P. et al. Ecology and economics for pandemic prevention. *Science* **369**, 379–381 (2020).
63. Di Marco, M. et al. Sustainable development must account for pandemic risk. *Proc. Natl. Acad. Sci. USA* **117**, 3888–3892 (2020).
64. Wang, H. et al. Scientific discovery in the age of artificial intelligence. *Nature* **620**, 47–60 (2023).
65. Sarker, I. H. Machine learning: algorithms, real-world applications and research directions. *SN Comput. Sci.* **2**, 160 (2021).
66. Li, H. et al. Diversity and transmission and zoonotic potential of microbes in true insectivores. *Nat. Commun.* **16**, 6709 (2025).
67. Franklins, L. H. V., Jones, K. E., Redding, D. W. & Abubakar, I. The effect of global change on mosquito-borne disease. *Lancet Infect. Dis.* **19**, e302–e312 (2019).
68. Tsui, J. L.-H. et al. Toward optimal disease surveillance with graph-based active learning. *Proc. Natl. Acad. Sci. USA* **121**, (2024).
69. Koh, P. W. & Liang, P. Understanding black-box predictions via influence functions. Preprint at <https://doi.org/10.48550/ARXIV.1703.04730> (2017).
70. Yuan, H., Yu, H., Gui, S. & Ji, S. Explainability in graph neural networks: a taxonomic survey. *IEEE Trans. Pattern Anal. Mach. Intell.* 1–19 (2022).
71. Díez-Villaseñor, C. & Rodríguez-Valera, F. CRISPR analysis suggests that small circular single-stranded DNA smacoviruses infect Archaea instead of humans. *Nat. Commun.* **10**, 294 (2019).
72. IUCN. *The IUCN Red List of Threatened Species* (IUCN, 2020).
73. Winter, D. J. rentrez: an R package for the NCBI eUtils API. Preprint at <https://doi.org/10.7287/peerj.preprints.3179v2> (2017).
74. Wickham, H. rvest: easily harvest (scrape) web pages <https://rvest.tidyverse.org/>, <https://github.com/tidyverse/rvest> (2025).
75. Wickham, H., Hester, J. & Ooms, J. xml2: Parse XML. R package version 1.6.0 <https://github.com/r-lib/xml2> (2026).
76. 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).
77. Faurby, S. et al. PHYLACINE 1.2: the phylogenetic atlas of mammal macroecology. *Ecology* **99**, 2626–2626 (2018).
78. Downs, C. J., Dochtermann, N. A., Ball, R., Klasing, K. C. & Martin, L. B. The effects of body mass on immune cell concentrations of mammals. *Am. Nat.* **195**, 107–114 (2020).
79. Sheldon, B. C. & Verhulst, S. Ecological immunology: costly parasite defences and trade-offs in evolutionary ecology. *Trends Ecol. Evol.* **11**, 317–321 (1996).
80. Cable, J. M., Enquist, B. J. & Moses, M. E. The allometry of host–pathogen interactions. *PLoS ONE* **2**, e1130 (2007).
81. Baltimore, D. Expression of animal virus genomes. *Bacteriol. Rev.* **35**, 235–241 (1971).
82. Sanjuán, R., Nebot, M. R., Chirico, N., Mansky, L. M. & Belshaw, R. Viral mutation rates. *J. Virol.* **84**, 9733–9748 (2010).
83. Mahy, B. W. J. The evolution and emergence of RNA viruses. *Emerg. Infect. Dis.* **16**, 899–899 (2010).
84. Pybus, O. G., Tatem, A. J. & Lemey, P. Virus evolution and transmission in an ever more connected world. *Proc. R. Soc. B Biol. Sci.* **282**, 20142878 (2015).
85. Coffin, J. M. Structure and classification of retroviruses. In *The Retroviridae* (ed. Levy, J. A.) 19–49 (Springer US, 1992).
86. Nisole, S. & Saïb, A. Early steps of retrovirus replicative cycle. *Retrovirology* **1**, 9 (2004).
87. Wawrzyniak, P., Płucienniczak, G. & Bartosik, D. The different faces of rolling-circle replication and its multifunctional initiator proteins. *Front. Microbiol.* **8**, 2353 (2017).
88. Lin, X. et al. Order and disorder control the functional rearrangement of influenza hemagglutinin. *Proc. Natl. Acad. Sci. USA* **111**, 12049–12054 (2014).
89. Lowen, A. C. Constraints, drivers, and implications of Influenza A virus reassortment. *Annu. Rev. Virol.* **4**, 105–121 (2017).
90. 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).
91. Villemereuil, P. D., Wells, J. A., Edwards, R. D. & Blomberg, S. P. Bayesian models for comparative analysis integrating phylogenetic uncertainty. *BMC Evol. Biol.* **12**, 102 (2012).
92. Faurby, S. & Svenning, J.-C. A species-level phylogeny of all extant and late Quaternary extinct mammals using a novel heuristic-hierarchical Bayesian approach. *Mol. Phylogenet. Evol.* **84**, 14–26 (2015).
93. Pebesma, E. Simple features for R: standardized support for spatial vector data. *R. J.* **10**, 439 (2018).
94. Van Dam, A. et al. Correspondence analysis, spectral clustering and graph embedding: applications to ecology and economic complexity. *Sci. Rep.* **11**, 8926 (2021).
95. Longdon, B., Brockhurst, M. A., Russell, C. A., Welch, J. J. & Jiggins, F. M. The evolution and genetics of virus host shifts. *PLoS Pathog.* **10**, e1004395 (2014).
96. Wardeh, M., Risley, C., McIntyre, M. K., Setzkorn, C. & Baylis, M. Database of host–pathogen and related species interactions, and their global distribution. *Sci. Data* **2**, 150049 (2015).
97. Wells, K., Morand, S., Wardeh, M. & Baylis, M. Distinct spread of DNA and RNA viruses among mammals amid prominent role of domestic species. *Glob. Ecol. Biogeogr.* **29**, 470–481 (2020).
98. Allen, T. et al. Global hotspots and correlates of emerging zoonotic diseases. *Nat. Commun.* **8**, 1124 (2017).

99. Dutilh, B. E., Reyes, A., Hall, R. J. & Whiteson, K. L. Editorial: Virus discovery by metagenomics: the (im)possibilities. *Front. Microbiol.* **8**, 1710 (2017).
100. Bompiani, E., Petrillo, U. F., Jona Lasinio, G. & Palini, F. High-performance computing with TeraStat. In *2020 IEEE International Conference on Dependable, Autonomic and Secure Computing, International Conference on Pervasive Intelligence and Computing, International Conference on Cloud and Big Data Computing, International Conference on Cyber Science and Technology Congress (DASC/PiCom/CBDCom/CyberSciTech)* 499–506 (IEEE, 2020).
101. Plummer, M. JAGS: a program for analysis of Bayesian graphical models using Gibbs sampling. In *Proc. 3rd International Workshop on Distributed Statistical Computing (DSC 2003)*, Vienna, Austria (2003).
102. Python. *Python Software Foundation* (Python, 2023).
103. Miasko, T. & Nowotny, M. PyJAGS: the Python interface to JAGS <http://github.com/michaelnowotny/pyjags> (2022).
104. Vehtari, A., Gelman, A., Simpson, D., Carpenter, B. & Bürkner, P.-C. Rank-normalization, folding, and localization: an improved R^2 for assessing convergence of MCMC (with discussion). *Bayesian Anal.* **16**, (2021).
105. Li, G., Xiong, C., Thabet, A. & Ghanem, B. DeeperGCN: all you need to train deeper GCNs. Preprint at <https://doi.org/10.48550/ARXIV.2006.07739> (2020).
106. He, K., Zhang, X., Ren, S. & Sun, J. Delving deep into rectifiers: surpassing human-level performance on ImageNet classification. Preprint at <https://doi.org/10.48550/ARXIV.1502.01852> (2015).
107. Cai, T. et al. GraphNorm: a principled approach to accelerating graph neural network training. Preprint at <https://doi.org/10.48550/ARXIV.2009.03294> (2020).
108. Paszke, A. et al. PyTorch: an imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems* Vol. 32 8024–8035 (Curran Associates, Inc., 2019).
109. Fey, M. & Lenssen, J. E. Fast graph representation learning with PyTorch geometric. In *ICLR Workshop on Representation Learning on Graphs and Manifolds* (2019).
110. Kingma, D. P. & Ba, J. Adam: a method for stochastic optimization. Preprint at <https://doi.org/10.48550/ARXIV.1412.6980> (2014).
111. Loshchilov, I. & Hutter, F. Decoupled weight decay regularization. Preprint at <https://doi.org/10.48550/ARXIV.1711.05101> (2017).
112. Kiryo, R., Niu, G., Plessis, M. C. du & Sugiyama, M. Positive-unlabeled learning with non-negative risk estimator. Preprint at <https://doi.org/10.48550/ARXIV.1703.00593> (2017).
113. Micikevicius, P. et al. Mixed precision training. Preprint at <https://doi.org/10.48550/arXiv.1710.03740> (2018).
114. Saito, Y., Yaginuma, S., Nishino, Y., Sakata, H. & Nakata, K. Unbiased recommender learning from missing-not-at-random implicit feedback. In *Proc. 13th International Conference on Web Search and Data Mining* 501–509 (ACM, 2020).
115. Allouche, O., Tsoar, A. & Kadmon, R. Assessing the accuracy of species distribution models: prevalence, kappa and the true skill statistic (TSS). *J. Appl. Ecol.* **43**, 1223–1232 (2006).
116. *Proc. 20th International Conference on Machine Learning*, August 21–24, 2003, Washington, DC, USA (AAAI Press, 2003).
117. Pignalberi, G., Tonelli, A., Giagu, S. & Di Marco, M. Replication data and codes for ‘Predicting unknown mammalian viral hosts using Dynamic Positive-Unlabeled learning. *Zenodo* <https://doi.org/10.5281/ZENODO.20764782> (2026).

Author contributions

G.P., A.T., S.G., and M.D.M. developed the study design. G.P. developed the methodological framework, curated the codes, conducted models training, generated predictions, and performed the analyses. G.P. and A.T. compiled the data. A.T. curated the generation of synthetic data. S.G. supervised the methodological framework. A.T. and M.D.M. provided ecological and eco-epidemiological expertise. G.P. and M.D.M. wrote the first draft of the manuscript. All authors discussed the results and improved the writing of the paper.

Funding

G.P. and S.G. disclose support for the research for this work from PNRR MUR project PE0000013-FAIR. M.D.M. discloses support for the research for this work from the EU project INF-ACT under the Next-Generation EU-MUR PNRR Mission 4 Component 2 Investment 1.3 (project PE00000007, INF-ACT Spoke4). A.T. declares no relevant funding.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41467-026-76416-4>.

Correspondence and requests for materials should be addressed to Gabriele Pignalberi or Moreno Di Marco.

Peer review information *Nature Communications* thanks the anonymous reviewer(s) for their contribution to the peer review of this work. A peer review file is available.

Reprints and permissions information is available at <http://www.nature.com/reprints>

Publisher’s note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article’s Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2026

This Article in Press is shared early to give you faster access to new research. It is citable and carries a permanent DOI. The final edited version will replace it automatically.