bioRxiv preprint doi: https://doi.org/10.1101/2020.05.22.111344; this version posted June 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Predicting wildlife hosts of betacoronaviruses for SARS-CoV-2 sampling prioritization

Daniel J. Becker[1,†], Gregory F. Albery[2,†], Anna R. Sjodin[3], Timothée Poisot[4], Tad A. Dallas[5], Evan A. Eskew[6,7], Maxwell J. Farrell[8], Sarah Guth[9], Barbara A. Han[10], Nancy B. Simmons[11], Michiel Stock[12], Emma C. Teeling[13], and Colin J. Carlson[14,15,*]

† These authors share lead author status

* Corresponding author: colin.carlson@georgetown.edu

1. Department of Biology, Indiana University, Bloomington, IN, U.S.A.
2. Department of Biology, Georgetown University, Washington, D.C., U.S.A.
3. Department of Biological Sciences, University of Idaho, Moscow, ID, U.S.A.
4. Université de Montréal, Département de Sciences Biologiques, Montréal, QC, Canada.
5. Department of Biological Sciences, Louisiana State University, Baton Rouge, LA, U.S.A.
6. Department of Ecology, Evolution, and Natural Resources, Rutgers University, New Brunswick, NJ, U.S.A.
7. Department of Biology, Pacific Lutheran University, Tacoma, WA, U.S.A.
8. Department of Ecology & Evolutionary Biology, University of Toronto, Toronto, ON, Canada.
9. Department of Integrative Biology, University of California Berkeley, Berkeley, CA, U.S.A.
10. Cary Institute of Ecosystem Studies, Millbrook, NY, U.S.A.
11. Department of Mammalogy, Division of Vertebrate Zoology, American Museum of Natural History, New York, NY, U.S.A.
12. KERMIT, Department of Data Analysis and Mathematical Modelling, Ghent University, Belgium
13. School of Biology and Environmental Science, Science Centre West, University College Dublin, Belfield, Dublin, Ireland.
14. Center for Global Health Science and Security, Georgetown University Medical Center, Washington, D.C., U.S.A.
15. Department of Microbiology and Immunology, Georgetown University Medical Center, Washington, D.C., U.S.A.

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.22.111344; this version posted June 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Abstract.

Despite massive investment in research on reservoirs of emerging pathogens, it remains difficult to rapidly identify the wildlife origins of novel zoonotic viruses. Viral surveillance is costly but rarely optimized using model-guided prioritization strategies, and predictions from a single model may be highly uncertain. Here, we generate an ensemble of eight network- and trait-based statistical models that predict mammal-virus associations, and we use model predictions to develop a set of priority recommendations for sampling potential bat reservoirs and intermediate hosts for SARS-CoV-2 and related betacoronaviruses. We find over 200 bat species globally could be undetected hosts of betacoronaviruses. Although over a dozen species of Asian horseshoe bats (*Rhinolophus* spp.) are known to harbor SARS-like coronaviruses, we find at least two thirds of betacoronavirus reservoirs in this bat genus might still be undetected. Although identification of other probable mammal reservoirs is likely beyond existing predictive capacity, some of our findings are surprisingly plausible; for example, several civet and pangolin species were highlighted as high-priority species for viral sampling. Our results should not be over-interpreted as novel information about the plausibility or likelihood of SARS-CoV-2's ultimate origin, but rather these predictions could help guide sampling for novel potentially zoonotic viruses; immunological research to characterize key receptors (e.g., ACE2) and identify mechanisms of viral tolerance; and experimental infections to quantify competence of suspected host species.

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.22.111344; this version posted June 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Main text.

Coronaviruses are a diverse family of positive-sense, single-stranded RNA viruses, found widely in mammals and birds[1]. They have a broad host range, a high mutation rate, and the largest genomes of any RNA viruses, but they have also evolved mechanisms for RNA proofreading and repair, which help to mitigate the deleterious effects of a high recombination rate acting over a large genome[2]. Consequently, coronaviruses fit the profile of viruses with high zoonotic potential. There are seven human coronaviruses (two in the genus *Alphacoronavirus* and five in *Betacoronavirus*), of which three are highly pathogenic in humans: SARS-CoV, SARS-CoV-2, and MERS-CoV. These three are zoonotic and widely agreed to have evolutionary origins in bats[3–6].

Our collective experience with both SARS-CoV and MERS-CoV illustrate the difficulty of tracing specific animal hosts of emerging coronaviruses. During the 2002–2003 SARS epidemic, SARS-CoV was traced to the masked palm civet (*Paguma larvata*)[7], but the ultimate origin remained unknown for several years. Horseshoe bats (family Rhinolophidae: *Rhinolophus*) were implicated as reservoir hosts in 2005, but their SARS-like viruses were not identical to circulating human strains[4]. Stronger evidence from 2017 placed the most likely evolutionary origin of SARS-CoV in *Rhinolophus ferrumequinum* or potentially *R. sinicus*[8]. Presently, there is even less certainty in the origins of MERS-CoV, although spillover to humans occurs relatively often through contact with dromedary camels (*Camelus dromedarius*). A virus with 100% nucleotide identity in a ~200 base pair region of the polymerase gene was detected in *Taphozous* bats (family Emballonuridae) in Saudi Arabia[9]; however, based on spike gene similarity, other sources treat HKU4 virus from *Tylonycteris* bats (family Vespertilionidae) in China as the closest-related bat virus[10,11]. Several bat coronaviruses have shown close relation to MERS-CoV, with a surprisingly broad geographic distribution from Mexico to China[12,13,14,15].

Coronavirus disease 2019 (COVID-19) is caused by severe acute respiratory syndrome coronavirus-2 (SARS-CoV-2), a novel virus with presumed evolutionary origins in bats. Although the earliest cases were linked to a wildlife market, contact tracing was limited, and there has been no definitive identification of the wildlife contact that resulted in spillover nor a true “index case.” Two bat viruses are closely related to SARS-CoV-2: RaTG13 bat CoV from *Rhinolophus affinis* (96% identical overall), and RmYN02 bat CoV from *Rhinolophus malayanus* (97% identical in one gene but only 61% in the receptor-binding domain and with less overall similarity)[6,16]. The divergence time between these bat viruses and human SARS-CoV-2 has been estimated as 40-50 years[17], suggesting that the main host(s) involved in spillover remain unknown. Evidence of viral recombination in pangolins has been proposed but is unresolved[17]. SARS-like betacoronaviruses have been recently isolated from Malayan pangolins (*Manis javanica*) traded in wildlife markets[18,19], and these viruses have a very high amino acid identity to SARS-CoV-2, but only show a ~90% nucleotide identity with SARS-CoV-2 or Bat-CoV RaTG13[20]. None of these host species are universally accepted as the origin of SARS-CoV-2 or a progenitor virus, and a “better fit” wildlife reservoir could likely still be identified. However, substantial gaps in betacoronavirus sampling

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.22.111344; this version posted June 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

across wildlife limit actionable inference about plausible reservoirs and intermediate hosts for SARS-CoV-2[21].

Identifying likely reservoirs of zoonotic pathogens is challenging[22]. Sampling wildlife for the presence of active or previous infection (i.e., seropositivity) represents the first stage of a pipeline for proper inference of host species[23], but sampling is often limited in phylogenetic, temporal, and spatial scale by logistical constraints[24]. Given such restrictions, modeling efforts can play a critical role in helping to prioritize pathogen surveillance by narrowing the set of plausible sampling targets[25]. For example, machine learning approaches have generated candidate lists of likely, but unsampled, primate reservoirs for Zika virus, bat reservoirs for filoviruses, and avian reservoirs for *Borrelia burgdorferi*[26–28]. In some contexts, models may be more useful for identifying which host or pathogen groups are *unlikely* to have zoonotic potential[29]. However, these approaches are generally applied individually to generate predictions. Implementation of multiple modeling approaches collaboratively and simultaneously could reduce redundancy and apparent disagreement at the earliest stages of pathogen tracing and help advance modeling work by addressing inter-model reliability, predictive accuracy, and the broader utility (or inefficacy) of such models in zoonosis research.

Because SARS-like coronaviruses (subgenus *Sarbecovirus*) are only characterized from a small number of bat species in publicly available data, current modeling methods are poorly tailored to exactly infer their potential reservoir hosts. In this study, we instead conduct two predictive efforts that may help guide the inevitable search for known and future zoonotic coronaviruses in wildlife: (1) broadly identifying bats and other mammals that may host any *Betacoronavirus* and (2) specifically identifying species with a high viral sharing probability with the two *Rhinolophus* species carrying the closest known wildlife relatives of SARS-CoV-2. To do this, we developed a standardized dataset of mammal-virus associations by integrating a previously published mammal-virus dataset[30] with a targeted scrape of all GenBank accessions for Coronaviridae and their associated hosts. Our final dataset spanned 710 host species and 359 virus genera, including 107 mammal hosts of betacoronaviruses as well as hundreds of other (non-coronavirus) association records. We integrated our host-virus data with a mammal phylogenetic supertree[31] and over 60 ecological traits of bat species[27,32,33]. We used these standardized data to generate eight predictive models of host-virus associations, including four network-based approaches, three trait-based approaches, and one hybrid approach using network and trait information. These efforts generated eight ranked lists of suspected bat hosts of betacoronaviruses and five ranked lists for other mammals. Each ranked list was scaled proportionally and consolidated in an ensemble of recommendations for betacoronavirus sampling and broader eco-evolutionary research (Supplemental Figure 1).

In our ensemble, we draw on two popular approaches to identify candidate reservoirs and intermediate hosts of betacoronaviruses. *Network-based methods* estimate a full set of "true" unobserved host-virus interactions based on a recorded network of associations (here, pairs of host species and associated viral genera). These methods are increasingly popular as a way to

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.22.111344; this version posted June 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

identify latent processes structuring ecological networks[34–36], but they are often confounded by sampling bias and can only make predictions for species within the observed network (i.e., those that have available virus data; in-sample prediction). In contrast, *trait-based methods* use observed relationships concerning host traits to identify species that fit the morphological, ecological, and/or phylogenetic profile of known host species of a given pathogen and rank the suitability of unknown hosts based on these trait profiles[28,37]. These methods may be more likely to recapitulate patterns in observed host-pathogen association data (e.g., geographic biases in sampling, phylogenetic similarity in host morphology), but they more easily correct for sampling bias and can predict host species without known viral associations (out-of-sample prediction).

Predictions of bat betacoronavirus hosts derived from network- and trait-based approaches displayed strong inter-model agreement within-group, but less with each other (Figure 1A,B). In-sample, we identified bat species across a range of genera as having the highest predicted probabilities of hosting betacoronaviruses, distributed in distinct families in both the Old World (e.g., Hipposideridae, several subfamilies in the Vespertilionidae) and the New World (e.g., *Artibeus jamaicensis* from the Phyllostomidae; Figure 1C). Out-of-sample, our multi-model ensemble more conservatively limited predictions to primarily Old World families such as Rhinolophidae and Pteropodidae (Figure 1D). Of the 1,037 bat hosts not currently known to host betacoronaviruses, our models identified between 1 and 720 potential hosts based on a 10% omission threshold (90% sensitivity). Applying this same threshold to our ensemble predictions, we identified 239 bat species that are likely undetected hosts of betacoronaviruses. These include more half of bat species in the genus *Rhinolophus* not currently known to be betacoronavirus hosts (35 of 61), compared to 16 known hosts in this genus. Given known roles of rhinolophids as hosts of SARS-like coronaviruses, our results suggest that SARS-like coronavirus diversity could be undescribed for around two-thirds of the potential reservoir bat species.

Our multi-model ensemble predicted undiscovered betacoronavirus bat hosts with striking geographic patterning (Figure 2). In-sample, the top 50 predicted bat hosts were broadly distributed and recapitulated observed patterns of bat betacoronavirus hosts in Europe, parts of sub-Saharan Africa, and Southeast Asia, although our models also predicted greater-than-expected richness of likely bat reservoirs in the Neotropics and North America. In contrast, the top out-of-sample predictions clustered in Vietnam, Myanmar, and southern China.

Because only trait-based models were capable of out-of-sample prediction, the differences in geographic patterns of our predictions likely reflect distinctions between the network- and trait-based modeling approaches, which we suggest should be considered qualitatively different lines of evidence. Network approaches proportionally upweight species with high observed viral diversity, recapitulating sampling biases largely unrelated to coronaviruses (e.g., frequent screening for rabies lyssaviruses in vampire bats, which have been sampled in a comparatively limited capacity for coronaviruses[14,38–40]). Highly ranked species may also have been previously sampled without evidence of betacoronavirus presence; for example, *Rhinolophus luctus* and

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.22.111344; this version posted June 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Macroglossus sobrinus from China and Thailand, respectively, tested negative for betacoronaviruses, but detection probability was limited by small sample sizes[41–43]. In contrast, trait-based approaches are constrained by their reliance on phylogeny and ecological traits, and the use of geographic covariates made models more likely to recapitulate existing spatial patterns of betacoronavirus detection (i.e., clustering in southeast Asia). However, their out-of-sample predictions are, by definition, inclusive of unsampled hosts[44], which potentially offer greater return on viral discovery investment.

Multi-model ensemble predictions also clustered taxonomically along parallel lines. Applying a graph partitioning algorithm (phylogenetic factorization) to the bat phylogeny[45], we found that in-sample predictions were on average lowest for the Yangochiroptera (Figure 3). This makes intuitive sense, because this clade does not include the groups known to harbor the majority of betacoronaviruses detected in bats (e.g., *Rhinolophus*, Hipposideridae). Out-of-sample predictions were lower in the New World superfamily Noctilionoidea, the emballonurids, and the *Lasiurus* genus, whereas the *Rhinolophus* genus and the Old World fruit bats (Pteropodidae) both had higher mean probabilities of betacoronavirus hosting (Supplemental Table 1).

These clade-specific patterns of predicted probabilities across extant bats could be particularly applicable for guiding future surveillance. On the one hand, betacoronavirus sampling in southeast Asian bat taxa (especially the genus *Rhinolophus*) may have a high success of viral detection but may not improve existing bat sampling gaps[46]. On the other hand, discovery of novel betacoronaviruses in Neotropical bats or Old World fruit bats could significantly revise our understanding of the bat-virus association network. Such discoveries would be particularly important for global health security, given the surprising identification of a MERS-like virus in Mexican bats[14] and the likelihood that post-COVID pandemic preparedness efforts will focus disproportionately on Asia despite the near-global presence of bat betacoronaviruses.

Although our ensemble model of potential bat betacoronavirus reservoirs generated strong and actionable predictions, our mammal-wide predictions were largely uninformative. In particular, minimal inter-model agreement (Supplemental Figure 2) indicated a lack of consistent, biologically meaningful findings. Major effects of sampling bias were apparent from the top-ranked species, which were primarily domestic animals or well-studied mesocarnivores (ED Figure 2B). Phylogenetic factorization mostly failed to find specific patterns in prediction (Supplemental Table 2): in-sample, mean predictions primarily confirmed betacoronavirus detection in the remaining Laurasiatheria (e.g., ungulates, carnivores, pangolins, hedgehogs, shrews), although nested clades of marine mammals (i.e., cetaceans) were less likely to harbor these viruses. Although cetaceans are predicted to be susceptible to SARS-CoV-2 based on ACE2 similarity with humans[47], this result is expected given betacoronavirus epidemiology and their predominance in terrestrial mammals. Our mammal predictions thus reflect a combination of detection bias and poor performance of network methods on limited data that likely signals the limits of existing predictive capacity. Our dataset contained only 30 non-bat betacoronavirus hosts, many of which were identified during sampling efforts following the first SARS outbreak[7].

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.22.111344; this version posted June 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Although the laurasiatherians include more potential intermediate hosts than other mammals, likely driven by inclusion of bats in this group, the high species diversity of this clade restricts insights for sampling prioritization, experimental work, or spillover risk management.

Given the unresolved origins of SARS-CoV-2 and significant motivation to identify other SARS-like coronaviruses and their reservoir hosts for pandemic preparedness[21], we further explored our only model that could generate out-of-sample predictions for all mammals[48]. This model uses geographic distributions and phylogenetic relatedness to estimate viral sharing probability. Where one or more (potential) hosts are known, these sharing patterns can be interpreted to identify probable reservoir hosts[48]. Because *Rhinolophus affinis* and *R. malayanus* host viruses that are closely related to SARS-CoV-2[6,16], we used their predicted sharing patterns to identify possible reservoirs of sarbecoviruses. In doing so, we aimed to work around a major data limitation: fewer than 20 sarbecovirus hosts were recorded in our dataset, a sample size that would preclude most modeling approaches.

For both presumed bat host species of sarbecoviruses, the most probable viral sharing hosts were again within the Laurasiatheria. Although bats—especially rhinolophids—unsurprisingly assumed the top predictions given phylogenetic affinity with known hosts (Supplemental Table 3, Supplemental Figure 3), several notable patterns emerged in the rankings of other mammals. Pangolins (Pholidota) were disproportionately likely to share viruses with *R. affinis* and *R. malayanus* (Supplemental Figure 4); the Sunda pangolin (*Manis javanica*) and Chinese pangolin (*M. pentadactyla*) were in the top 20 predictions for both reservoir species (Supplemental Table 4). This result is promising given the much-discussed discovery of SARS-like betacoronaviruses in *M. javanica*[18]. The Viverridae were also disproportionately well-represented in the top predictions (Supplemental Figure 5), most notably the masked palm civet (*Paguma larvata*), which was identified as an intermediate host of SARS-CoV[49,50] (Supplemental Table 4). Our virus sharing model thus captured historic patterns of betacoronavirus spillover and predicted cross-species transmission from only the phylogeography of east Asian mammals. The opportunity for inter-species contact depends on both animal behavior and habitat selection, which could be better studied in natural and anthropogenic habitats for these clades[51]. However, over evolutionary timescales, our results suggest recombination and diversification of novel betacoronaviruses should be expected among such taxa in nature, without necessarily requiring further contact via the wildlife trade. For example, recent observations in Gabon demonstrate cohabitation between pangolins, bats, and various other mammals (e.g., rodents) in burrows[52].

Moreover, these findings lend credibility to other predictions of SARS-CoV-2 sharing patterns and host susceptibility. Many of the model's top predictions were mustelids (i.e., ferrets and weasels), and the most likely viral sharing partner for both *Rhinolophus* species was the hog badger (*Arctonyx collaris*; Supplemental Table 4). Taken together with reports of SARS-CoV-2 spread in mink farms[53], these results highlight the relatively unexplored potential for mustelids to serve as betacoronavirus hosts and as models for infectivity studies[54]. Similarly, identification of several deer and Old World monkey taxa as high-probability hosts in our clade-based analysis

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.22.111344; this version posted June 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

(Supplemental Figure 3) meshes with the observation of high binding of SARS-CoV-2 to ACE2 receptors in cervid deer and primates[47]. Felids (especially leopards) also ranked relatively high in our viral sharing predictions (Supplemental Table 4, Supplemental Figure 5), which is of particular interest given reports of SARS-CoV-2 susceptibility among cats[55]. However, we caution that this model was the only approach in our ensemble that could generate out-of-sample prediction across mammals, and therefore its predictions lacked confirmation (and filtering of potential spurious results) by other models that were designed and implemented independently.

Several limitations apply to our work, most notably the difficulty of empirically verifying predictions. Although some virological studies have incidentally tested specific hypotheses (e.g., filovirus models and bat surveys[27,56], henipavirus models and experimental infections[23,57]), model-based predictions are nearly never subject to systematic verification or post-hoc efforts to identify and correct spurious results. Greater dialogue between modelers and empiricists is necessary to systematically confront the growing set of predicted host-virus associations with experimental validation or field observation.

Already, we have identified roughly a half-dozen new bat reservoirs of betacoronaviruses that our study correctly predicted. *Scotophilus heathii*, *Hipposideros larvatus*, and *Pteropus lylei*, all highly predicted bat species in our out-of-sample rankings, have been reported positive for betacoronaviruses in the literature[43,58]; however, resulting sequences were not annotated to genus level in GenBank. More recently, the release of 630 novel bat coronaviruses sequences from China included four additional bat hosts (*Hipposideros pomona, Scotophilus kuhlii, Myotis pequinius,* and *M. horsfieldii*)[59]. All but *M. horsfieldii* were correctly identified by the ensemble, and three models correctly identified all seven verification hosts (Trait-based 1 and 3 and Network-based 1). These results support the idea that our models identified relevant targets correctly but also highlight an evident limitation of the workflow. Whereas an automated approach was the ideal method to systematically compile over 30,000 samples on the timescales commensurate with ongoing efforts to trace SARS-CoV-2 in wildlife, we suggest this discrepancy highlights the need for careful virological work downstream at every stage of the modeling process, including the development of hybrid manual-automated data pipelines.

Additionally, overcoming underlying model biases that are driven by historical sampling regimes will require coordinated efforts in field study design. Bat sampling for betacoronaviruses has prioritized viral discovery[39,40,60–62], but limitations in the spatial and temporal scale (and replication) of field sampling have likely created fundamental gaps in our understanding of infection dynamics in bat populations[24]. Limited longitudinal sampling of wild bats suggests betacoronavirus detection is sporadic over time and space[58,63], implying strong seasonality in virus shedding pulses[64]. Carefully tailored spatial and temporal sampling efforts for priority taxa identified here, within the *Rhinolophus* genus or other high-prediction bat clades, will be key to identifying the environmental drivers of betacoronavirus shedding from wild bats and possible opportunities for contact between bats, intermediate hosts, and humans.

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.22.111344; this version posted June 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Future field studies will undoubtedly be important to understand viral dynamics in bats but are inherently costly and labor-intensive. These efforts are particularly challenging during a pandemic, as many scientific operations have been restricted including field studies of bats in some regions to limit possible viral spillback from humans. However, various alternative efforts could both advance basic virology and allow testing model predictions. General open access to viral association records, including GenBank accessions and the upcoming release of the USAID PREDICT program's data, could answer open questions and allow updates to our sampling prioritization (including potentially modeling at subgenus level, with greater data availability). Appropriately preserved museum specimens and historical collections, from diverse research programs, also offer key opportunities to retrospectively screen samples from bats and other mammals for betacoronaviruses and to enhance our understanding of complex host-virus interactions[65]. Large-scale research networks, such as GBatNet (Global Union of Bat Diversity Networks; https://gbatnet.blogspot.com) and its member networks, could provide diverse samples and ensure proper partnerships and equitable access and benefit sharing of knowledge across countries[66,67]. Whole-genome sequencing through initiatives such as the Bat1K Project (https://bat1k.ucd.ie) would facilitate fundamental and applied insights into the immunological pathways by which bats can harbor many virulent viruses without displaying disease[68,69]. Additionally, targeted sequencing efforts could also identify endogenized viral elements in bat genomes, shedding light on the diversity of bat viruses and the evolution of bat immune systems[70,71].

To expedite such work, we have made our binary predictions of host-virus associations for all eight models and all 1,000+ bat species publicly available (Supplementary File 1). Such results are provided both in the spirit of open science and with the hope that future viral detection, isolation, or experimental studies might confirm some of these predictions or rule out others[57]. In ongoing collaborative efforts, we aim to consolidate results from field studies that address these predictions (e.g., serosurveys) and to track Genbank submissions to expand the known list of betacoronavirus hosts. In several years, we intend to revisit these predictions as a post-hoc test of model validation, which would represent the first effort to test the performance of such models and assess their contribution to basic science and to pandemic preparedness.

It is crucial that our predictions be interpreted as a set of hypotheses about potential host-virus compatibility rather than strong evidence that a particular mammal species is a true reservoir for betacoronaviruses. In particular, susceptibility is only one aspect of host competence[22,72], which encompasses the diverse genetic and immunological processes that mediate within-host responses following exposure[73]. SARS-CoV-2 in particular may have a broad host range[47], given hypothesized compatibility with the ACE2 receptor in many mammal species, but this only adds to the extreme caution with which any data should be used to implicate a potential wildlife reservoir of the virus, given that rapid interpretation of inconclusive molecular evidence has likely already generated spurious reservoir identifications[74,75]. Future efforts to isolate live virus from wildlife or to experimentally show viral replication would more robustly test whether predicted host species actually play a role in betacoronavirus maintenance in wildlife[57].

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.22.111344; this version posted June 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Without direct lines of virological evidence, we note that our sampling prioritization scheme also does not implicate any given mammal species in SARS-CoV-2 transmission to humans. Care should be taken to communicate this, especially given the potential consequences of miscommunication for wildlife conservation. The bat research community in particular has expressed concern that negative framing of bats as the source of SARS-CoV-2 will impact public and governmental attitudes toward bat conservation[76]. In zoonotic virus research on bats, studies often over-emphasize human disease risks[77] and rarely mention ecosystem services provided by these animals [78]. Skewed communication can fuel negative responses against bats, including indiscriminate culling (i.e., reduction of populations by selective slaughter)[79], which has already occured in response to COVID-19 even outside of Asia (where spillover occurred)[80].

To minimize potential unintended negative impacts for bat conservation, public health and conservation responses should act in accordance with substantial evidence suggesting that culling has numerous negative consequences, not only threatening population viability of threatened bat species in shared roosts[81] but also possibly increasing viral transmission within the very species that are targeted[82,83]. Instead, bat conservation programs and long-term ecological studies are necessary to help researchers understand viral ecology and find sustainable solutions for humans to live safely with wildlife. From another perspective, policy solutions aimed at limiting human-animal contact could potentially prevent virus establishment in novel species (e.g., as observed in mink farms[53]), especially in wildlife that may already face conservation challenges (e.g., North American bats threatened by an emerging disease, white-nose syndrome[79,84]). At least two bat species that can be infected by the fungal pathogen (*Eptesicus fuscus* and *Tadarida brasiliensis*) are in our list of the 239 bat species most likely to be betacoronavirus hosts.

Substantial investments are already being planned to trace the wildlife origins of SARS-CoV-2. However, the intermediate progenitor virus may never be isolated from samples contemporaneous with spillover, and it may no longer be circulating in wildlife. MERS-CoV circulates continuously in camels[85] and SARS-CoV persisted in civets long enough to seed secondary outbreaks[49,50], but the limited description of Pangolin-CoV symptoms suggests high mortality, potentially indicating a more transient epizootic such as Ebola die-offs in red river hogs (*Potamochoerus porcus*)[18]. In lieu of concrete data, our study provides no additional evidence implicating any particular species—or any particular pathway of spillover (e.g., wildlife trade, consumption of hunted animals)—as more or less likely. No specific scenario can be confirmed or rigorously interrogated by ecological models, and we explicitly warn against misinterpretation or misuse of our findings as evidence for adjacent policy decisions. Although policies that focus on particular potential reservoir species or target human-wildlife contact could reduce future spillovers, they will have a negligible bearing on the ongoing pandemic, as SARS-CoV-2 is highly transmissible within humans (e.g., unlike MERS-CoV or other zoonoses that are sustained in people by constant reintroduction). SARS-CoV-2 is likely to remain circulating in human populations at least until a vaccine is developed (if not much longer, given current challenges to

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.22.111344; this version posted June 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

386 vaccine production, allocation, governance, and uptake), regardless of immediate actions
387 regarding wildlife. COVID-19 response must be informed by the best consensus evidence
388 available and prioritize solutions that address immediate reduction of transmission through
389 public health and policy channels. Meanwhile, we hope our proposed wildlife sampling priorities
390 will help increase the odds of preventing the future emergence of novel betacoronaviruses.
391

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.22.111344; this version posted June 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

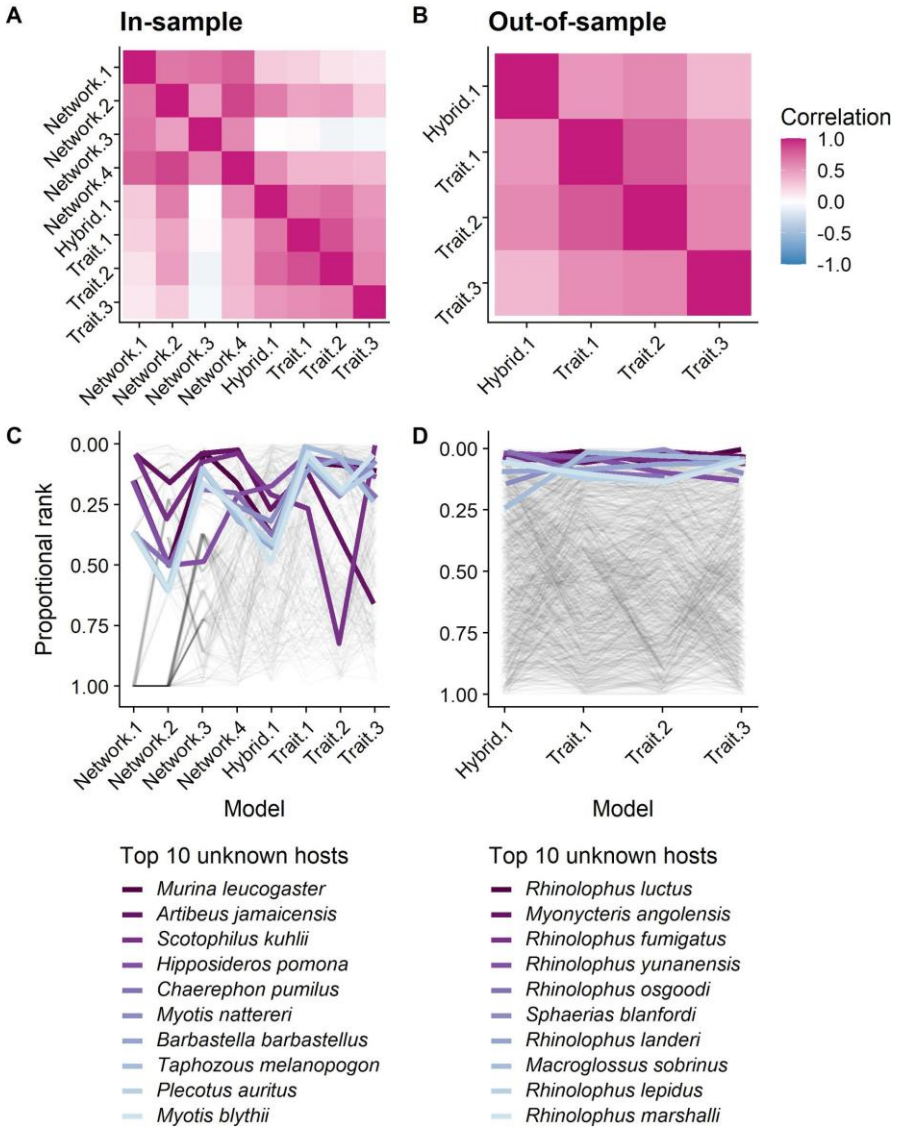
Acknowledgements

We thank Heather Wells for generously sharing thoughtful comments and code. The VERENA consortium is supported by L'Institut de Valorisation de Données (IVADO) through Université de Montreal. DJB was supported by an appointment to the Intelligence Community Postdoctoral Research Fellowship Program at Indiana University, administered by Oak Ridge Institute for Science and Education through an interagency agreement between the U.S. Department of Energy and the Office of the Director of National Intelligence. MS was supported by the Research Foundation - Flanders (FWO17/PDO/067) and the Flemish Government under the "Onderzoeksprogramma Artificiële Intelligentie (AI) Vlaanderen" program.

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.22.111344; this version posted June 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

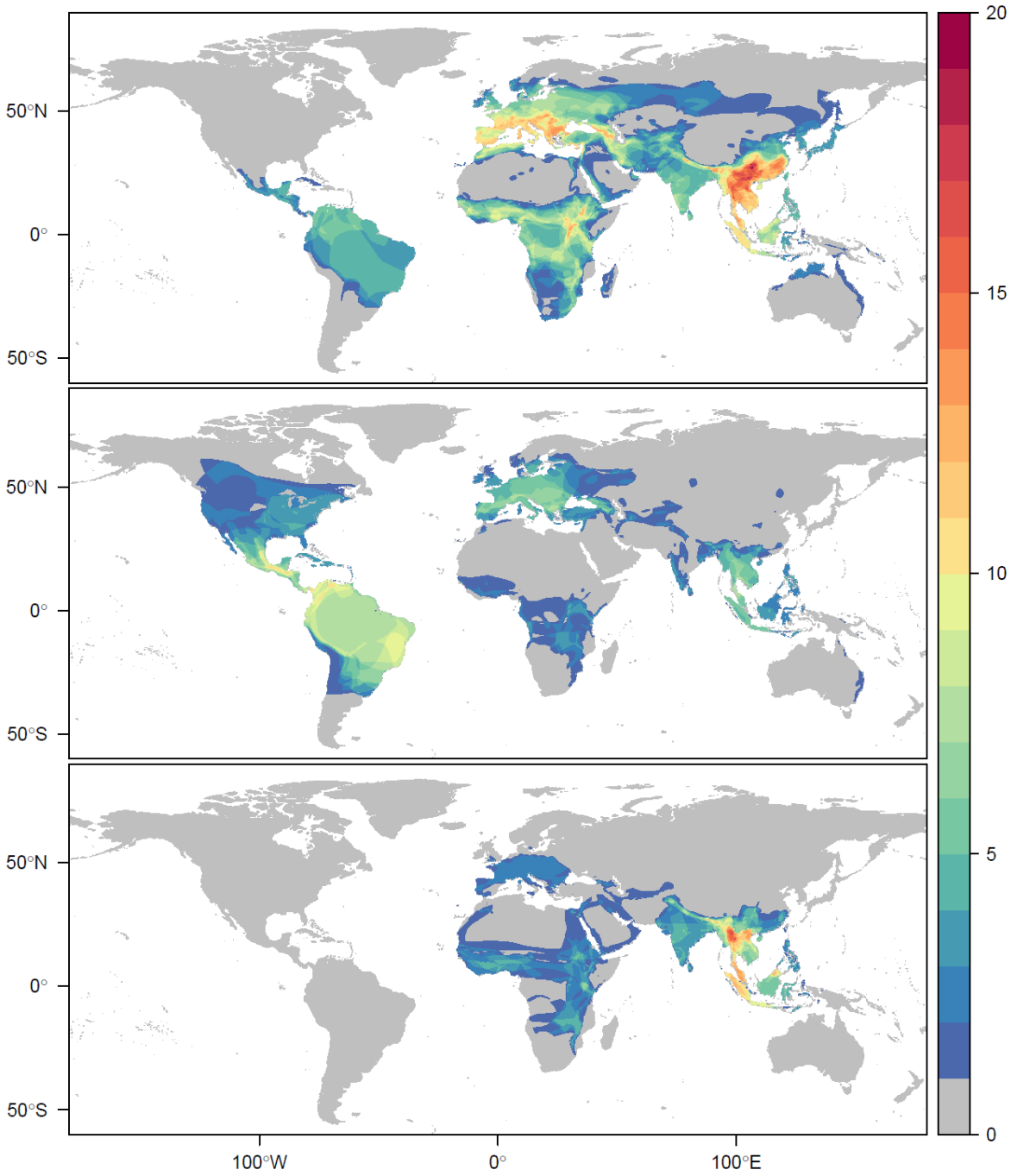
Figures

Figure 1. An ensemble of predictive models facilitates identification of likely betacoronavirus bat hosts. The pairwise Spearman's rank correlations between models' ranked species-level predictions were generally substantial and positive (A,B). Models are arranged in decreasing order of their mean correlation with other models. In-sample predictions, expressed as host species' proportional rank (0 is the most likely host from a given model, 1 is the least likely host), varied significantly due to the uncertainty of network approaches (C). In contrast, species' proportional ranks were tightly correlated across out-of-sample predictive approaches, which relied on species traits (D). Each line represents a different bat species' proportional rank across models. The ten species with the highest mean proportional ranks across all models are highlighted in shades of purple.



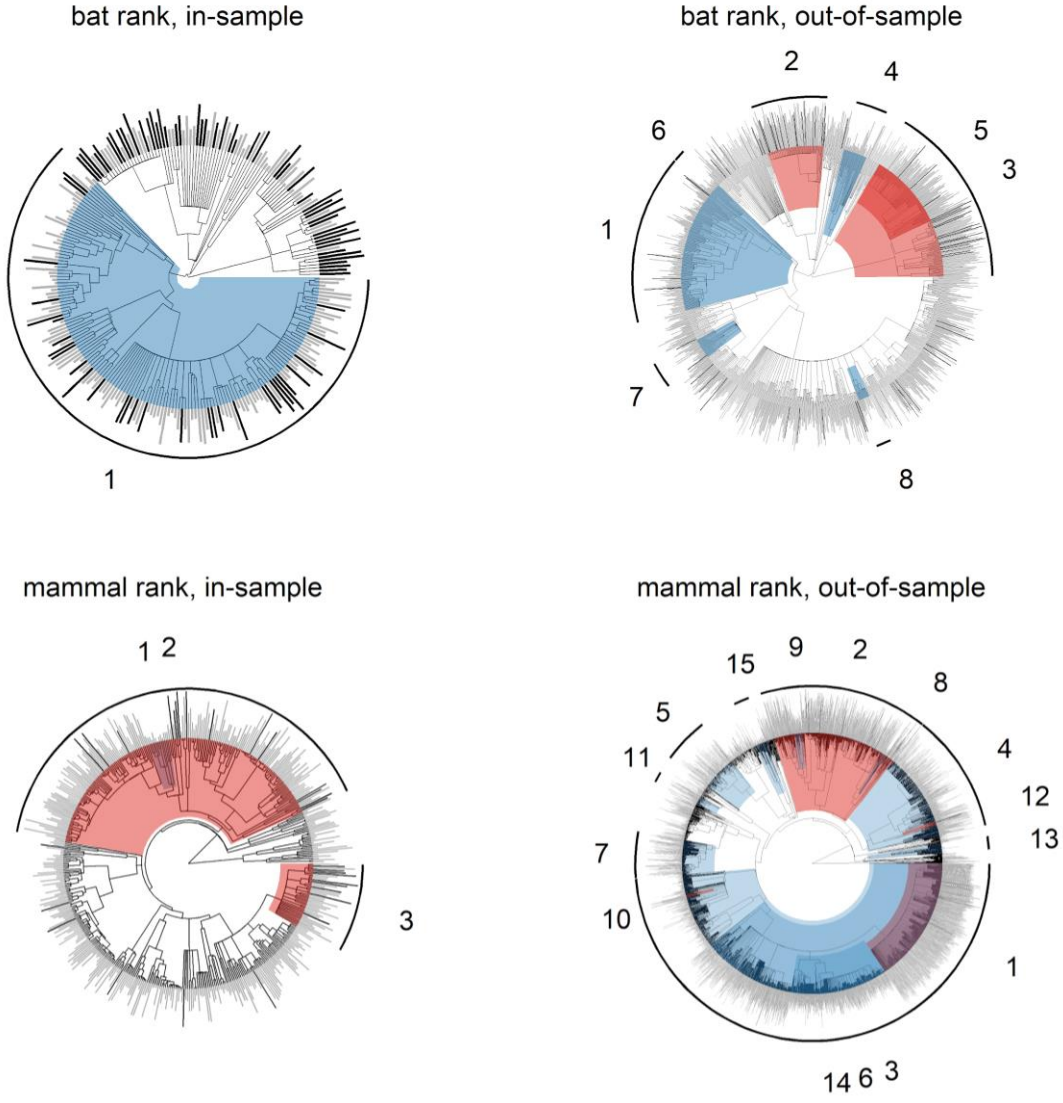
bioRxiv preprint doi: https://doi.org/10.1101/2020.05.22.111344; this version posted June 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Figure 2. Species richness of known and suspected betacoronavirus bat hosts. Known hosts of betacoronaviruses (*top*) are found worldwide, but particularly in southern Asia and southern Europe. The top 50 predicted bat hosts with viral association records (*middle*) are mostly Neotropical, including several species of vampire bats. In contrast, the top 50 *de novo* bat host predictions based on phylogeny and ecological traits (*bottom*) are mostly clustered in Myanmar, Vietnam, and southern China, with none in the Neotropics or North America.



bioRxiv preprint doi: https://doi.org/10.1101/2020.05.22.111344; this version posted June 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Figure 3. Phylogenetic distribution of predicted bat and mammal hosts of betacoronaviruses. Bar height indicates mean predicted rank across the model ensemble (higher values = lower proportional rank score, more likely to be a host) and black indicates known betacoronavirus hosts. Colored regions indicate clades identified by phylogenetic factorization as significantly different in their predicted rank compared to the paraphyletic remainder; those clades more likely to contain a host are shown in red, whereas those less likely to contain a host are shown in blue. Results are displayed for bats and all mammals separately, stratified by in- and out-of-sample predictions. Numbers reference clade names, species richness, and mean predicted ranks as described in Supplemental Tables 1 and 2.



bioRxiv preprint doi: https://doi.org/10.1101/2020.05.22.111344; this version posted June 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Methods.

The underlying conceptual aim of this study was to produce and synthesize several different models that predict and rank candidate reservoir species—each with different methods, assumptions, and framings—and to rapidly synthesize these into a consensus list. We broadly structured our study around two modeling targets: (1) produce rankings of likely bat hosts of betacoronaviruses and (2) identify potential non-bat mammal hosts. We developed a novel dataset that merged existing knowledge about the broader mammal-virus network with targeted data collection about coronaviruses; implemented eight modeling methods; synthesized these into an ensemble; and post-hoc identified taxonomic patterns in prediction using phylogenetic factorization.

Host-Virus Association Data

Entries were downloaded from GenBank on March 27th 2020 using the following search terms: Coronavirus, Coronaviridae, Orthocoronavirinae Alphacoronavirus, Betacoronavirus, Gammacoronavirus, and Deltacoronavirus. Data were sorted using a Python script that saved all available metadata regarding accession number, division, submission date, entry title, organism, genus, genome length, host classification, country, collection date, PubMed ID, journal containing associated publication, publication year, genome completeness, and the gene sequenced. The dataset was cleaned to remove duplicate entries, using GenBank accession number, and entries that did not correspond to viral sequences, using GenBank division. After cleaning, 31,473 entries remained, of which 25,628 had metadata regarding host species.

Data from GenBank were merged with the Host-Pathogen Phylogeny Project (HP3) dataset[30]. The HP3 dataset consists of 2,805 associations between 754 mammal hosts and 586 virus species, compiled from the International Committee on Taxonomy of Viruses (ICTV) database, and manually cleaned over a period of five years. Data collection on HP3 began in 2010 and has been static since 2017, but it still represents the most complete dataset on the mammal virome published with a high standard of data documentation. Several recent studies have used the HP3 dataset to produce statistical models of viral sharing or zoonotic potential[29,48,86], making it a comparable reference for a multi-model ensemble study.

Because of naming inconsistencies both within GenBank and between the two datasets (HP3 and GenBank), we used a two-step pipeline for taxonomic reconciliation. Viral names were matched to the ICTV 2019 master species list, up to the sub-genus level. Host species names were matched against GBIF using their species API with an automated Julia script, and processed to a fully cleaned set of names. This led to an harmonized dataset representing a global list of mammal-virus associations, from which the bat-coronavirus data can be extracted for downstream and specific modeling efforts. Because the HP3 dataset used an older version of the ICTV master list, and because not all host names in the GenBank metadata could be matched by

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.22.111344; this version posted June 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

the GBIF species API (or could be solved unambiguously to the species level), some host-virus interactions were lost; this reinforces the need to careful data curation of taxonomic metadata if they are to enable and support predictive pipelines.

Predictor Data

Phylogeny

We used a supertree of extant mammals to unify modeling approaches incorporating host phylogeny[31]. Although more recent mammal supertrees exist, we used this particular phylogeny for consistency with trait datasets and several of the modeling frameworks included in our ensemble. We manually matched select bat species names between our edge list and this particular phylogeny. This included reverting any *Dermanura* to their former *Artibeus* designation (i.e., *D. phaeotis*, *D. cinerea*, *D. tolteca*)[87], switching *Tadarida* species to either *Mops* or *Chaerophon* species (i.e., *Tadarida condylura* to *Mops condylurus*, *Tadarida plicata* to *Chaerephon plicatus*, *Tadarida pumila* to *Chaerephon pumilus*)[88], and renaming *Myotis pilosus* to the more recent *Myotis ricketti*. *Chaerephon pusillus* was considered its own species but is now synonymous with *Chaerephon pumilus*[88]. Minor discrepancies between virus data and our phylogeny were also corrected (*Hipposideros commersonii* to *Hipposideros commersoni* [although more recently changed to *Macronycteris commersoni*], *Rhinolophus hildebrandti* to *Rhinolophus hildebrandtii*, *Neoromicia nana* to *Neoromicia nanus*). In other cases, some recently revised genera in our edge list were modified to match former genera in the mammal supertree: *Parastrellus hesperus* to *Pipistrellus hesperus*, and *Perimyotis subflavus* to *Pipistrellus subflavus*[89]. Lastly, some names in our edge list missing from the mammal supertree represent former subspecies being raised to full species rank, and names were reverted accordingly: *Artibeus planirostris* to *Artibeus jamaicensis*, *Miniopterus fuliginosus* to *Miniopterus schreibersii*, *Triaenops afer* to *Triaenops persicus*, and *Carollia sowelli* to *Carollia brevicauda*. Although we recognize that these are each now recognized as distinct species, in all cases our synonymized names are thought to be either sister taxa or very closely related.

Ecological traits

We used a previously published dataset of 63 ecological traits describing the morphology, life history, biogeography, and diet of 1,116 bat species. These data are drawn from a combination of PanTHERIA[32], EltonTraits[33], and the IUCN Red List range maps, and were previously cleaned in a study producing predictions of bat reservoirs of filoviruses[27]. Four redundant variables (two for human population density, mean potential evapotranspiration in range, and body mass) were eliminated prior to analyses, favoring variables with higher completeness.

Correction for sampling bias

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.22.111344; this version posted June 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

To correct for sampling bias, in the style of several previous studies[30,86], we used the number of peer-reviewed citations available on a given host as a measure of scientific sampling effort. We used the R package *easyPubMed* to scrape the number of citations in PubMed returned when searching each of the 1,116 bat names in the trait data on April 10, 2020.

Modeling Approaches

Our team produced an ensemble of eight statistical models (Supplemental Tables 5 and 6), and applied them to generate a predictive set of eight models for bats and five for other mammals. Four use a network-theoretic component (k-nearest neighbors, linear filtering, trait-free plug-and-play, and scaled phylogeny), while three primarily used ecological traits as predictors (boosted regression trees, Bayesian additive regression trees, and neutral phylogeographic). A final hybrid model used a combination approach with a network backbone but incorporating the phylogeny and bat trait data.

All eight approaches were used to generate predictions about potential bat hosts of betacoronaviruses. A subset of five were used to recommend potential non-bat mammal hosts of betacoronaviruses (*k*-nearest neighbor, linear filtering, scaled phylogeny, trait-free plug-and-play, and neutral phylogeographic). We did not use trait-based models to predict non-bat hosts, because assigning pseudoabsences to the vast majority (~3500 or more) of mammal species would likely lead to largely uninformative predictions, weighed against the 109 known betacoronavirus hosts (79 bats and 30 other mammals).

Network model 1: k-nearest neighbors recommender

We follow the methodology previously developed for the recommendation of species feeding interactions[90]. This method builds a recommender system internally based on the *k*-nearest neighbor (kNN) algorithm, under which candidate hosts are recommended for a virus from a pool constituted by the hosts of the *k* viruses with which it has the greatest overlap. Overlap (host sharing) is measured using Tanimoto similarity, which is the cardinality of the intersection of two sets divided by the cardinality of their union. To obtain the pairwise similarity between two viruses, this divides the number of shared hosts by the cumulative number of hosts. The *k*NN of a virus are the *k* other viruses with which it has the highest Tanimoto similarity.

Hosts are then recommended by counting how many times they appear in these *k* neighbors, a quantity that ranges from 1 to *k*. We can impose arbitrary cutoffs by limiting the recommendations to the hosts that occur in at least *k*, *k-1*, etc, viruses. Previous leave-one-out validation of this model revealed that it is particularly effective for viruses with a reduced number of hosts, which is likely to be the case for emerging viruses. Furthermore, the performance of this model was not significantly improved by the addition of functional traits, making it acceptable to run on the association data only.

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.22.111344; this version posted June 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

This model was run two times: first, by measuring the similarity of viruses, and recommending hosts; second, by measuring the similarity of hosts, and recommending viruses. In all cases, only results for betacoronaviruses are reported.

The outcome of this model should be subject to caution, as leave-one-out validation revealed that the success rate (i.e., ability to recover one interaction that has been removed) remained lower than 50% even when using k=8, and dropped as low as 5% when using k=1 (the nearest-neighbor algorithm). This strongly suggests that the dataset of reported host-virus associations is extremely incomplete; therefore, the identification of the nearest neighbors can be biased by under-reported interactions, and this can result in noise in the prediction. This noise can be particularly important when the *k*NN technique operates on viruses, of which the bat dataset has only 15.

Network-based model 2: Linear filter recommender

Following Stock *et al.*[91], we used a previously developed linear filter to infer potential missing interactions. This recommender system assumes that networks tend to be self-similar, and use this information to generate a score for an un-observed interaction that is a linear combination of the status of the interaction (relative weight of 1/4), relative degree of host and virus, and of the observed connectance of the network (all with relative weights of 1); as we are concerned with ranking interactions as opposed to examining the absolute value of the score, the penalization coefficient associated to the interaction being presumed absent could be omitted with no change in the ranking, but has been set to a low value instead. The scores returned by the linear filter are not directly related to the probability of the interaction existing in this context, but higher scores still indicate interactions that are more likely to exist. Indeed, known hosts of betacoronavirus typically scored higher.

We used the zero-one-out approach to assess the performance of this model on the entire datasets. In all cases, non-interactions ranked lower than positive interactions even when entirely removing the penalization coefficient from the linear filter parameters, which suggests that the network structure (degree and connectance) is capturing a lot of information as to which species can interact.

Network-based model 3: Plug and play

For network problems, the “plug and play” model is a statistical approach that formulates Bayes’ theorem for link prediction around the conditional density of traits of known associations compared to traits of every possible association in a network. The conditional density function is measured by using non-parametric kernel density estimators (implemented with the R package *np*), and the conditional ratio between them is used to estimate link “suitability”, a scale-free ratio. Compared to other machine learning methods that fit training data iteratively, plug and play is comparatively simple, and directly infers the most likely extensions of observed patterns in data.

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.22.111344; this version posted June 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

The plug and play was originally developed to forecast missing links in host-parasite networks[36], but has since been used to model species distributions[92] and predict the global spread of human infectious diseases[93]. We used this model here to estimate suitability of host-virus interactions by first modeling the entire estimated network of host-virus interaction suitability, and ranking hosts that are not infected by betacoronaviruses by their estimated suitability for betacoronaviruses.

The "plug and play" model is trained using either matched pairs of host and pathogen ecological, morphological, or phylogenetic traits[36], or by using a latent approach[93] which considers the mean similarity of pathogens in their host ranges and the mean similarity of hosts in their pathogen communities as 'traits'. We decided to use the latent approach, as host trait data was far more available than viral trait data. Further, the taxonomic scale considered for host (species) and virus (genus) differed, making the resolution of potential trait data different enough to potentially confound trait-based approaches in this modeling framework.

Relative suitability of a host-virus association, as estimated by the "plug and play" model, is formulated as a density ratio estimation problem. The suitability of a host-virus association is quantified as the quotient of the distribution of latent trait values when an association was recorded over the distribution of all the latent trait values. As an attempt to control for sampling effort of mammal and bat host species, we included PubMed citation counts for host species (as described above) in the estimation of host-virus suitability. We explored host-pathogen suitability using the entire mammal-virus associations dataset, to maximize the available information on the network's structure, and ranked host-pathogen pairs by their relative suitability value. From the final predictions, we subset out bat-specific predictions. When predicting, we set citation counts to the mean of training data, as a sampling bias correction.

Network-based model 4: Scaled-phylogeny

We apply the network-based conditional model of Elmasri *et al.*[94] for predicting missing links in bipartite ecological networks. The full model combines a hierarchical Bayesian latent score framework which accounts for the number of interactions per taxon, and a dependency among hosts based on evolutionary distances. To predict links based on evolutionary distance, the probability of a host-parasite interaction is taken as the sum of evolutionary distances to the documented hosts of that parasite. This allocates higher probabilities when a few closely related hosts, or many distantly related hosts interact with a parasite. In this way phylogenetic distances are combined with individual affinity parameters per taxa to model the conditional probability of an interaction.

In ecological studies, it is common to use time-scaled phylogenies to quantify evolutionary distance among species[95]. We may use these fixed evolutionary distances for link prediction, but parasite taxa are known to be more or less constrained by phylogenetic distances among hosts[96]. Further, phylogenies are hypotheses about evolutionary relationships and have uncertainties in

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.22.111344; this version posted June 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

the topology and relative distances among species[97]. Rather than treating phylogenetic distances as fixed, Elmasri *et al.*[94] re-scale the phylogeny by applying a macroevolutionary model of trait evolution. While any evolutionary model that re-scales the covariance matrix may be used, we use the early-burst model, which allows evolutionary change to accelerate or decelerate through time[98]. This different emphasis to be placed on deep versus recent host divergences when predicting links.

We apply the model to a network of associations among host species and viral genera, and the mammal supertree, which allows us to leverage information from across the network to predict undocumented bat-betacoronavirus associations. We fit sets of models, applying both the full model, and the phylogeny-only model to both the bat-viral genera associations, and the mammal-viral genera associations. For each data-model combination we fit the model using ten-fold cross-validation holding out links for which there is a minimum of two observed interactions. The posterior interaction matrices resulting from each of the ten models are then averaged to generate predictions for all links in the network, with betacoronaviruses subset to generate the ensemble predictions.

To assess predictive performance, we attempted to predict the held out interactions, and calculated AUC scores by thresholding predicted probabilities per fold, and taking an average across the 10 folds. In addition to AUC, we also assessed the model based on the percent of documented interactions accurately recovered. For the bat-viral genera data the full model resulted in an average AUC of 0.82 and recovered an average of 90.1% of held out interactions, while the phylogeny-only model showed increased AUC (0.86), but a decreased proportion of held-out interactions recovered (84.5%). Interestingly, the models for bat-virus genera associations had marginally worse predictive performance compared to the same models run on the larger network of mammal-virus associations (full model: AUC 0.88, 84.4% positive interactions recovered; phylogeny-only model:AUC: 0.88, 88.8% positive interactions recovered), indicating that predicting bat-betacoronavirus associations may benefit from including data on non-bat hosts. The models also estimated the scaling parameter (eta) of the early-burst model to be positive (average eta=7.92 for the full model run on the bat subset), indicating accelerating evolution compared to the input tree (Supplemental Figure 6). This means that recent divergences are given more weight than deeper ones for determining bat-viral genera associations, which is consistent with recent work on viral sharing[48,99].

Trait-based model 1: Boosted regression trees

Previous work has been highly successful in predicting zoonotic reservoirs using a combination of taxonomic, ecological, and geographic traits as predictors. This approach has been previously used to identify wildlife hosts of filoviruses[27,100], flaviviruses[28,101], henipaviruses[23], *Borrelia burgdorferi*[26], to predict mosquito vectors of flaviviruses[102], and to predict rodent reservoirs and tick vectors of zoonotic viruses[37,103]. These approaches treat the presence of a specific virus (or genus of viruses) or a zoonotic pathogen as an outcome variable, with negative values given for

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.22.111344; this version posted June 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

690 species not known to be hosts (pseudoabsences), and use machine learning to identify the
691 characteristics that predispose animals to hosting pathogens of concern. By predicting the
692 probability that a given pseudoabsence is a false negative, the method can infer potential
693 undetected or undiscovered host species.
694
695 This approach has almost exclusively been implemented using boosted regression trees (BRT),
696 a classification and regression tree (CART) machine learning method that became popular a
697 decade ago for species distribution modeling.[104] Boosted regression trees develop an ensemble
698 of classification trees which iteratively explain the residuals of previous trees, up to a fixed tree
699 depth (usually between 3 and 5 splits). The incorporation of boosting allows the model, as it is fit,
700 to progressively better explain poorly-fit cases within training data.
701
702 We used boosted regression trees to identify trait profiles that predict bat hosts of
703 betacoronaviruses, including all trait predictors from the trait database that met baseline
704 coverage (< 50% missing values) and variation (< 97% homogenous) thresholds. For all model
705 fitting, we specified a Bernoulli error distribution for our binary response variable and applied 10-
706 fold cross validation to prevent overfitting (R package *gbm*). We started by fitting a global model
707 to our full dataset, first specifying learning rate = 0.01 (shrinks the contribution of each tree to the
708 model) and tree complexity = 4 (controls tree depth) as per default values and subsequently
709 tuning to minimize cross-validation error.
710
711 We reduced the variable set by calling the gbm.simp() function, which computes and compares
712 the mean change in cross-validation error (deviance) produced by dropping different sets of least-
713 contributing predictors. The final simplified model included 23 variables, plus citation counts,
714 which we added to correct for sampling bias.
715
716 We applied bootstrapping resampling methods to estimate uncertainty, using our tuned model to
717 fit 1000 replicate models. For each model, training sets were assembled by randomly selecting
718 with replacement 79 bat-coronavirus associations from the set of reported bat hosts and 79
719 pseudoabsences. Trained models were used to generate relative influence coefficients for trait
720 predictors and coronavirus host probabilities across all bat species. Partial dependence plots
721 display relative influence coefficients and bootstrapped confidence intervals for the top ten
722 contributing trait predictors. The medians of host probabilities were ranked and used to identify
723 the top ten candidate host species. When predicting, we set citation counts to the mean of training
724 data, as a sampling bias correction.
725
726 *Trait-based model 2: Bayesian additive regression trees*
727
728 A similar workflow to trait-based model 1 was implemented using Bayesian additive regression
729 trees (BART), an emerging machine learning tool that has similarities to more popular methods
730 like random forests and boosted regression trees. BART adds several layers of methodological
731 innovation, and performs well in bakeoffs with other advanced machine learning methods.

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.22.111344; this version posted June 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Several features make BART very convenient for modeling projects like these, including several easy-to-use implementations in R packages, built-in capacity to impute and predict on missing data, and easy construction of variable importance and partial dependence plots.

Like other classification and regression tree methods, BART assigns the probability of a binary outcome variable by developing a set of classification trees - in this case, a sum-of-trees model - that split data ("branches") and assign values to terminal nodes ("leaves"). Whereas other similar methods generate uncertainty by adjusting data (e.g. random forests bootstrap training data and fit a tree to each bootstrap; boosted regression trees are usually implemented with iterated training-test splits to generate confidence intervals), BART generates uncertainty using an MCMC process. An initial sum-of-trees model is fit to the entire dataset, and then rulesets are adjusted in a limited and stochastic set of ways (e.g., adding a split; switching two internal nodes), with the sum-of-trees model backfit to each change. After a burn-in period, the cumulative set of sum-of-trees models is treated as a posterior distribution. This has some advantages over other methods, like boosted regression trees or random forests. In particular, posterior width directly measures model uncertainty (rather than approximating it by permuting training data), and a single model can be run (instead of an ensemble trained on smaller subsets of training data), allowing the model to use the full training dataset all at once.[105]

Unlike many Bayesian machine learning methods, BART is easily implemented out-of-the-box, due to a limited set of customization needs. Three main priors control the fitting process: one usually-uniform prior on variable importance, one two-parameter negative power distribution on tree depth (preventing overfitting), and an inverse chi-squared distribution on residual variance. A set of well-performing priors from the original BART study[106] are widely used across R implementations for out-of-the-box settings, but can be further adjusted relative to modeling needs. In this study, we implemented BART models using a Dirichlet prior for variable importance (DART), a specification that is designed for situations with high dimensionality data that probably reflects a small number of true informative predictors. This often produces a much more reduced model without going through a stepwise variable selection process, which can be slow and very subject to stochasticity.[105]

We implemented this approach using the *BART* package in R, using the bat-virus association dataset to generate an outcome variable, and the bat traits dataset as predictors. BART models were implemented with 200 trees and 10,000 posterior draws, using every trait feature that was at least 50% complete and < 97% homogenous (taken from TBM1).

We tried four total implementations, based on two decisions: BART uncorrected and corrected for citation counts (BART-u, BART-c), and DART uncorrected and corrected for citation counts (DART-u, DART-c). All four models performed well, with little variation in predictive power measured by the area under the receiver operator curve calculated on training data (BART-u: AUC = 0.93; BART-c: AUC = 0.93; DART-u: AUC = 0.93; DART-c: 0.90; Supplemental Figure 7). Across all models, spatial variables had a high importance, including some regionalization (extent of

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.22.111344; this version posted June 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

774 range) and some variables capturing larger geographic range sizes, as did a diet of invertebrates
775 (pulling out the phylogenetic signal of insectivorous bats; Supplemental Figure 8).
776
777 All models identified a number of "false negative" hosts that would be suitable based on a 10%
778 false negative classification threshold for known betacoronavirus hosts (implemented with the R
779 package 'PresenceAbsence'). BART-u identified 217 missing hosts, BART-c identified 279
780 missing hosts, DART-u identified 222 missing hosts, and DART-c identified 384 missing hosts,
781 suggesting that this model most penalized overfitting as intended. As a result, we considered this
782 model the most rigorous and powerful for inference, and used DART-c in the final model
783 ensemble. We predicted across all 1,040 bats without recorded betacoronavirus associations,
784 and ranked predicted probability. When predicting, we set citation counts to the mean of training
785 data, as a sampling bias correction.
786
787 *Trait-based model 3: Phylogeographic neutral model*
788
789 We used a previously published pairwise viral sharing model[48] to predict potential
790 betacoronavirus hosts based on the sharing patterns of known hosts in a published dataset [30].
791 We used a generalised additive mixed model (GAMM), which was fitted in the first half of 2019
792 using the *mgcv* package, with pairwise binary viral sharing (0/1 denoting if a species shares at
793 least one virus) as a response variable. Explanatory variables include pairwise proportional
794 phylogenetic distance and geographic range overlap (taken from the IUCN species ranges), with
795 a multi-membership random effect to control for species-level sampling biases. The model was
796 then used to predict the probability that a given species pair share at least one virus across 4196
797 placental mammals with available data, producing a predicted viral sharing network that
798 recapitulates a number of known macroecological patterns, as well as predicting reservoir hosts
799 with surprising accuracy[48]. Subsetting this predicted sharing matrix, we listed the rank order of
800 hosts most likely to share with all known betacoronavirus hosts in our datasets.
801
802 *Rhinolophus-specific implementation of trait-based model 3*
803
804 We then repeated this process with sharing patterns of *Rhinolophus affinis* and *R. malayanus*
805 specifically. Given the strong phylogenetic effect, the top 139 predictions were bat species:
806 predominantly rhinolophids and hipposiderids. The top 20 predictions for both *R. malayanus* and
807 *R. affinis* are displayed in Supplemental Tables 3 and 4. Notable predictions included the hog
808 badger *Arctonyx collaris* (Carnivora: Mustelidae), which was examined for SARS-CoV antibodies
809 in 2003 and is reported in wildlife markets[7,107]; a selection of civet cats (Carnivora: Viverridae)
810 including *Viverra* species*;* the binturong (*Arctitis binturong*)*;* and the masked palm civet (*Paguma*
811 *larvata*), the latter of which were implicated in the chain of emergence for SARS-CoV[49,50]; and
812 pangolins (Pholidota: Manidae) including *Manis javanica* and *Manis pentadactyla,* which have
813 been hypothesised to be part of the emergence chain for SARS-CoV-2[18,19].
814

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.22.111344; this version posted June 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Alongside these high-ranked species-level predictions, we visually examined how predictions varied across all mammal orders and families using the whole dataset (Supplemental Figure 5). Pangolins (Pholidota), treeshrews (Scandentia), carnivores (Carnivora), hedgehogs (Erinaceomorpha), and even-toed ungulates (Artiodactyla) had high mean predicted probabilities. Investigating family-level sharing probabilities revealed that civets (Viverridae) and mustelids (Mustelidae) were responsible for the high Carnivora probabilities, and mouse deer (Tragulidae) and bovids (Bovidae) were mainly responsible for high probabilities in the Artiodactyla (Supplemental Figure 5).

Hybrid model 1: Two-Step Kernel Ridge Regression

To make predictions using both the traits and the phylogeny of the hosts, we used Two-Step Kernel Ridge Regression (TSKRR), a kernel-based pairwise learning method[108,109]. This method can predict interactions between two species, i.e., hosts and viruses, both in-sample and out-of-sample. Conceptually, TSKRR performs nonlinear regression twice: once to generalize to new hosts and once to generalize to new viruses (Supplemental Figure 9A). As a kernel method, it can exploit arbitrary similarity measures to describe the species. At the same time, its conceptual simplicity permits using efficient shortcuts for tuning and specialized cross-validation.

For the hosts, we again restricted ourselves to bats. First, we inputted missing trait values using the *MissForest* package[110]. We then selected the relevant traits, standardized them, and computed the squared Euclidean distance. We plugged this distance into a standard radial basis kernel[111], where we used the median heuristic to set the bandwidth. Similarly, for the bat phylogeny, we computed the distance matrix. We combined this distance with the radial basis kernel, again setting the bandwidth by the median of the distances. Bats were described using a kernel matrix derived from traits, phylogeny, or the average of both (Supplemental Figure 9B). We corrected for sampling bias by including the number of citations per species as a covariate and predicted using the average number of citations for all hosts. For viruses, we used a simple kernel with a “1” if two virus genera belong to the same family, and a “0” elsewise. For all the kernel matrices, we added 0.1 to each element and an additional 0.1 to the diagonal elements. This modification serves as an intercept and it gives the TSKRR the same capacity as the linear filtering method described earlier[108].

We fitted the models using the R package *xnet*. We assessed model performance in two ways. First, we used leave-one-out for every element of the incidence matrix and computed the AUC over all predicted interactions. Secondly, we used leave-one-out cross-validation on bats(i.e., leaving out a single species, making predictions for all the viruses of that bat, and repeating this for every species). In this setting, we computed the AUC for every virus over all bats and averaged these values (Supplemental Figure 9C). For both evaluation strategies, we removed viruses with no hosts in the dataset, although we kept these for training the model (Supplementary Table 7). Compared to the average performance for the viruses, betacoronaviruses were slightly more difficult to predict. Phylogeny was slightly more informative than traits, although the effect

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.22.111344; this version posted June 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

between them was synergistic. As would be expected, it was easier to predict missing interactions than to predict new hosts.

To assess which variables are most important, we computed feature importance based on random permutations. We randomly reshuffled one or both of the bats' kernel matrices, destroying its information content. Subsequently, we recorded the decrease in AUC for the setting 'interactions' in the model as an indication of relative importance. We also randomly reshuffled each trait separately and monitored the effect on performance. We repeated each random permutation 100 times. Reshuffling the complete kernel matrix with both traits and phylogeny resulted in the largest average drop in AUC of 0.1904 (Supplementary Figure 10). Reshuffling only the phylogeny kernel matrix had a much more profound impact than the trait matrix (0.07769 vs. 0.05301). Contributions of traits were smaller by comparison. Citation count was the largest effect (0.02119), with diet breadth (0.00520), population group size (0.003802), and annual birth pulse (0.0022944) as the remaining largest contributions. Removing citations resulted in a stronger importance for these variables but did not change the relative importance.

Consensus Methods and Recommendations

Combining and ranking predictions

For all eight models predicting bat hosts of betacoronaviruses, and five models predicting mammal hosts of betacoronaviruses, we combined predictions—generated using the same standardized data—into one standardized dataset. All mammal models were trained on data including bats, but predictions were subset to exclude bats to focus on likely intermediate hosts.

Each study's unique output—a non-intercomparable mix of different definitions of suitability or probability of association—were transformed into proportional rank, where lower rank indicates higher evidence for association out of the total number of hosts examined. By rescaling all results to proportional ranks between zero and one, we also allowed comparison of in-sample and out-of-sample predictions across all models. Proportional ranks were averaged across models to generate one standardized list of predictions. This absorbed much of the variation in model performance (Supplemental Figure 1) and produced a set of rankings that performed well.

We elected not to withhold any "test" data to measure model performance, given that each method deployed in the ensemble has been independently and rigorously tested and validated in previous publications. Instead, to maximize the amount of available training data for every model, we used full datasets in each model and measured performance on the full training data.

For bats, the final ensemble of models spanned a large range of performance on the training data, measured by the area under the receiver operator curve (AUC; Network 1: 0.624; Network 2: 0.987; Network 3: 0.514; Network 4: 0.726; Trait 1: 0.850; Trait 2: 0.902; Trait 3: 0.762; Hybrid 1: 0.924), indicating that it was possible to suitably detect differences in model performance on the full

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.22.111344; this version posted June 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

data. The total ensemble of proportional ranks performed medium well (AUC = 0.832). We used known betacoronavirus associations to threshold each model and the ensemble predictions based on a 10% omission threshold (90% sensitivity), and we again found a wide range in the number of predicted undiscovered bat hosts of betacoronaviruses (Network 1: 162 species; Network 2: 1; Network 3: 111; Network 4: 44; Trait 1: 425; Trait 2: 384; Trait 3: 720; Hybrid 1: 181; total ensemble: 239 species). Given concerns about mammal model performance and biological accuracy (see Main Text), we elected not to apply this exercise to mammal hosts at large.

To visualize the spatial distribution of predicted bat hosts, we used the IUCN Red List database of species geographic distributions. We took the top 50 ranked in-sample predictions and top 50 ranked out-of-sample predictions and combined these range maps to visualize species richness of top predicted hosts (Figure 3).

Phylogenetic factorization of ensemble models

We used phylogenetic factorization to flexibly identify taxonomic patterns in the consensus proportional rankings of likely hosts of SARS-CoV-2. Phylogenetic factorization is a graph-partitioning algorithm that iteratively partitions a phylogeny in a series of generalized linear models to identify clades at any taxonomic level (e.g., rather than *a priori* comparing strictly among genera or family) that differ in a trait of interest [45]. Using the mammal supertree, we used the *phylofactor* package to partition proportional rank as a Gaussian-distributed variable. We determined the number of significant phylogenetic factors using a Holm's sequentially rejective 5% cutoff for the family-wise error rate. We applied this algorithm across our four final ensemble prediction datasets: in-sample bat ranks, out-of-sample bat ranks, in-sample mammal ranks, and out-of-sample mammal ranks.

Using network and trait-based models within-sample, we identified only one bat clade with substantially different consensus proportional rankings, the Yangochiroptera ($\bar{x}$=0.55 compared to 0.39 for the remaining bat phylogeny, the Yinpterochiroptera). Out of sample, using only trait-based models, we instead identified eight bat clades with different propensities to include unlikely or likely bat hosts of betacoronaviruses. Subclades of the New World superfamily Noctilionoidea broadly had higher proportional ranks ($\bar{x}$=0.73), indicating lower predicted probability of being hosts, as did the Emballanuridae ($\bar{x}$=0.75). In contrast, both the Rhinolophidae and the Pteropodidae broadly had lower mean ranks ($\bar{x}$=0.29 and $\bar{x}$=0.38).

Using network models within-sample across non-volant mammals, we identified four clades with different proportional ranks. The largest clade was the Laurasiatheria (Artiodactyla, Perissodactyla, Carnivora, Pholidota, Soricomorpha, and Erinaceomorpha), which had lower proportional ranks (higher risk; $\bar{x}$=0.55). Nested within this clade, the Cetacea had greater proportional ranks ($\bar{x}$=0.89), indicating lower risk. A large subclade of the Murinae (Old World rats and mice) also had lower ranks ($\bar{x}$=0.52). Out of sample, using only the biogeographic viral sharing model, we instead identified 15 clades with different proportional ranks. The first clade identified large swaths of the Muridae as having higher risk ($\bar{x}$=0.38) as well as the Laurasiatheria ($\bar{x}$=0.50).

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.22.111344; this version posted June 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Old World primates had weakly lower risk ($\bar{x}$=0.65), as did the Scuridae ($\bar{x}$=0.67). The Cetacea and Pinnipedia both had greater proportional ranks ($\bar{x}$=0.89 and $\bar{x}$=0.71). Old World porcupines (Hystricidae) and the Erinaceidae (Paraechinus, Hemiechinus, Mesechinus, Erinaceus, Atelerix) both had greater risk ($\bar{x}$=0.48 and $\bar{x}$=0.39), while the Afrosoricida had higher ranks ($\bar{x}$=0.97).

To assess potential discrepancy between taxonomic patterns in model ensemble predictions and those of simply host betacoronavirus status itself, we ran a secondary phylogenetic factorization treating host status as a Bernoulli-distributed variable, with the same procedure applied to determine the number of significant phylogenetic factors. To assess sensitivity of taxonomic patterns to sampling effort, we ran phylogenetic factorization with and without square-root transformed PubMed citations per species as a weighting variable (Supplemental Figure 11).

Without accounting for study effort, phylogenetic factorization of betacoronavirus host status identified one significant clade across the bat phylogeny, the Yangochiroptera, as having fewer positive species (4.71%) than the paraphyletic remainder (12.12%). When accounting for study effort, however, the single clade identified by phylogenetic factorization changed, with a subclade of the family Pteropodidae (the Rousettinae) having a greater proportion of positive species (28.6%). For non-volant mammals, phylogenetic factorization identified only one clade, the family Camelidae, as having more positive species (75%) than the tree remainder (0.68%).

Phylogenetic factorization of Rhinolophidae virus sharing

Because phylogenetic patterns in predictions from our viral sharing model could vary across other taxonomic scales beyond order and family, we also used phylogenetic factorization to more flexibly identify host clades with different propensities to share viruses with *R. affinis* and *R. malayanus*. We partitioned rank as a Gaussian-distributed variable and again determined the number of significant phylogenetic factors using Holm's sequentially rejective 5% cutoff.

Within the Chiroptera, we identified 10 clades with different propensities to share viruses with *R. affinis* and 5 clades with different propensities to share viruses with *R. malayanus*. For both bats, the top clade was the family Rhinolophidae, reinforcing phylogenetic components of the biogeographic model and highlighting the greater likelihood of viral sharing (mean rank $\bar{x}$=40 for *R. affinis*, $\bar{x}$=42 for *R. malayanus*). For *R. affinis*, several individual bat species had lower risks of viral sharing (e.g., *Myotis leibii*, $\bar{x}$=4100; *Pteropus insularis*, $\bar{x}$=3157; *Nyctimene aello*, $\bar{x}$=2497; *Chaerephon chapini*, $\bar{x}$=2497). The Megadermatidae, Nycteridae, and Hipposideridae (under which the PanTHERIA dataset includes the genus *Rhinonicteris*, although this is now considered a separate family, the Rhinonycteridae[112]) collectively had greater likelihood of viral sharing ($\bar{x}$=557), as did the Vespertillionidae ($\bar{x}$=704).

Across the non-volant mammals, we identified 7 clades with different propensities to share viruses with *R. affinis* and only 1 clade with different propensities to share viruses with *R. malayanus*. For both bat species, the first and primary clade was the Ferungulata (Artiodactyla,

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.22.111344; this version posted June 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Perissodactyla, Carnivora, Pholidota, Soricomorpha, and Erinaceomorpha), which had lower ranks (higher viral sharing; $\bar{x}$=2084). For viral sharing with *R. affinis*, the Sciuridae was more likely to share viruses ($\bar{x}$=1948), as was the Scandentia ($\bar{x}$=1416) and many members of the Colobinae ($\bar{x}$=1958). However, members of the tribe Muntiacini (genera *Elaphodus* and *Muntiacus*) had especially high likelihoods of viral sharing and low rank ($\bar{x}$=361).

Data and Code Availability

The standardized data on betacoronavirus associations, and all associated predictor data, is available from the VERENA consortium's Github (github.com/viralemergence/virionette). All modeling teams contributed an individual repository with their methods, which are available in the organizational directory (github.com/viralemergence). All code for analysis, and a working reproduction of each authors' contributions, is available from the study repository (github.com/viralemergence/Fresnel).

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.22.111344; this version posted June 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Supplemental Material

Supplemental Table 1. Results of phylogenetic factorization applied to predicted rank probabilities for bats. The number of retained phylogenetic factors (following a 5% family-wise error rate applied to GLMs), taxa corresponding to those clades, number of species per clade, and mean predicted rank probabilities for the clade compared to the paraphyletic remainder are shown stratified by models applied in- and out-of-sample.

Sample	Factor	Taxa	Tips	Clade	Other
in	1	Yangochiroptera	160	0.550	0.393
out	1	Mystacinidae, Noctilionidae, Mormoopidae, Phyllostomidae	161	0.730	0.483
out	2	Rhinolophidae	73	0.295	0.538
out	3	Pteropodidae	173	0.379	0.548
out	4	Mosia, Emballonura, Coleura, Rhynchonycteris, Cyttarops, Diclidurus, Centronycteris, Cormura, Saccopteryx, Balantiopteryx, Peropteryx	31	0.755	0.512
out	5	Eonycteris, Macroglossus, Syconycteris, Notopteris, Melonycteris, Harpyionycteris, Aproteles, Dobsonia, Styloctenium, Neopteryx, Pteralopex, Acerodon, Pteropus	99	0.458	0.527
out	6	Thyropteridae, Furipteridae, Natalidae	12	0.818	0.517
out	7	Otomops, Promops, Molossus, Eumops	26	0.669	0.517
out	8	Lasiurus	15	0.717	0.518

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.22.111344; this version posted June 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Supplemental Table 2. Results of phylogenetic factorization applied to predicted rank probabilities for all mammals. The number of retained phylogenetic factors (following a 5% family-wise error rate applied to GLMs), taxa corresponding to those clades, number of species per clade, and mean predicted rank probabilities for the clade compared to the paraphyletic remainder are shown stratified by models applied in- and out-of-sample.

Sample	Factor	Taxa	Tips	Clade	Other
in	1	Phocoenidae, Delphinidae, Tursiops, Monodontidae, Physeteridae, Balaenopteridae, Eschrichtiidae	12	0.889	0.611
in	2	Artiodactyla, Perissodactyla, Carnivora, Pholidota, Erinaceomorpha, Soricomorpha	173	0.549	0.661
in	3	Lophuromys, Micaelamys, Apodemus, Arvicanthis, Bandicota, Madromys, Dasymys, Hydromys, Lemniscomys, Mastomys, Mus, Pelomys, Niviventer, Otomys, Praomys, Rattus, Vandeleuria	38	0.520	0.627
out	1	Abditomys, Bullimus, Limnomys, Tarsomys, Tryphomys, Acomys, Lophuromys, Uranomys, Aethomys, Micaelamys, Anisomys, Chiruromys, Coccymys, Crossomys, Hyomys, Leptomys, Lorentzimys, Pseudohydromys, Paraleptomys, Macruromys, Mallomys, Microhydromys, Parahydromys, Pogonomelomys, Abeomelomys, Solomys, Xenuromys, Apodemus, Tokudaia, Apomys, Crunomys, Chrotomys, Rhynchomys, Arvicanthis, Bandicota, Batomys, Carpomys, Crateromys, Berylmys, Bunomys, Chiromyscus, Chiropodomys, Hapalomys, Haeromys, Colomys, Nilopegamys, Conilurus, Leporillus, Mesembriomys, Melomys, Protochromys, Mammelomys, Paramelomys, Uromys, Zyzomys, Leggadina, Notomys, Pseudomys, Mastacomys, Madromys, Cremnomys, Millardia, Dacnomys, Dasymys, Dephomys, Hybomys, Hydromys, Xeromys, Desmomys, Diomys, Diplothrix, Echiothrix, Margaretamys, Melasmothrix, Tateomys, Eropeplus, Lenomys, Golunda, Grammomys, Thallomys, Hadromys, Heimyscus, Hylomyscus, Komodomys, Papagomys, Oenomys, Thamnomys, Lemniscomys, Lenothrix, Leopoldamys, Malacomys, Praomys, Myomyscus, Mastomys, Maxomys, Micromys, Muriculus, Mus, Mylomys, Pelomys, Stenocephalemys, Nesokia, Niviventer, Otomys, Parotomys, Palawanomys, Paruromys, Phloeomys, Pithecheir, Pogonomys, Rattus, Rhabdomys, Srilankamys, Nesoromys, Stochomys, Sundamys, Taeromys, Vandeleuria, Vernaya, Zelotomys	510	0.382	0.672
out	2	Artiodactyla, Perissodactyla, Carnivora, Pholidota	505	0.495	0.651
out	3	Anomaluridae, Pedetidae, Dipodidae, Cricetidae, Muridae, Nesomyidae, Calomyscidae, Spalacidae, Platacanthomyidae	779	0.643	0.622
out	4	Talpidae, Erinaceomorpha, Soricidae	357	0.630	0.627
out	5	Cercopithecidae, Hominidae, Hylobatidae	139	0.649	0.626

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.22.111344; this version posted June 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Supplemental Table 2, continued. (Page 2 of 2)

Sample	Factor	Taxa	Tips	Clade	Other
out	6	Abrawayaomys, Handleyomys, Aepeomys, Thomasomys, Abrothrix, Akodon, Necromys, Deltamys, Thaptomys, Andalgalomys, Auliscomys, Loxodontomys, Phyllotis, Paralomys, Graomys, Andinomys, Bibimys, Kunsia, Scapteromys, Blarinomys, Calomys, Chelemys, Chilomys, Chinchillula, Delomys, Eligmodontia, Euneomys, Galenomys, Geoxus, Holochilus, Lundomys, Pseudoryzomys, Irenomys, Lenoxus, Melanomys, Microryzomys, Neacomys, Nectomys, Neotomys, Nesoryzomys, Notiomys, Oecomys, Oligoryzomys, Oryzomys, Oxymycterus, Brucepattersonius, Phaenomys, Podoxymys, Punomys, Reithrodon, Rhagomys, Rhipidomys, Scolomys, Sigmodontomys, Thalpomys, Wiedomys, Wilfredomys, Juliomys, Zygodontomys, Anotomys, Chibchanomys, Ichthyomys, Neusticomys, Rheomys, Sigmodon, Nyctomys, Otonyctomys, Ototylomys, Tylomys, Baiomys, Scotinomys, Ochrotomys, Habromys, Neotomodon, Podomys, Osgoodomys, Megadontomys, Peromyscus, Onychomys, Isthmomys, Reithrodontomys, Hodomys, Xenomys, Neotoma, Nelsonia	397	0.703	0.616
out	7	Tamiasciurus, Sciurus, Rheithrosciurus, Microsciurus, Syntheosciurus, Pteromys, Petaurista, Belomys, Biswamoyopterus, Trogopterus, Pteromyscus, Aeromys, Eupetaurus, Aeretes, Glaucomys, Eoglaucomys, Hylopetes, Petinomys, Petaurillus, Iomys, Ratufa, Callosciurus, Glyphotes, Lariscus, Menetes, Rhinosciurus, Funambulus, Tamiops, Dremomys, Exilisciurus, Hyosciurus, Prosciurillus, Rubrisciurus, Nannosciurus, Sundasciurus	139	0.672	0.625
out	8	Phocoenidae, Delphinidae, Tursiops, Monodontidae, Physeteridae, Balaenopteridae, Eschrichtiidae	12	0.889	0.626
out	9	Odobenidae, Otariidae, Phocidae	33	0.714	0.626
out	10	Hystricidae	11	0.482	0.627
out	11	Caprolagus, Poelagus, Lepus, Oryctolagus	33	0.642	0.627
out	12	Paraechinus, Hemiechinus, Mesechinus, Erinaceus, Atelerix	15	0.388	0.628
out	13	Afrosoricida	41	0.970	0.623
out	14	Castoridae, Heteromyidae, Geomyidae, Octodontidae, Ctenodactylidae, Ctenomyidae, Abrocomidae, Caviidae, Dinomyidae, Petromuridae, Dasyproctidae, Myocastoridae, Echimyidae, Erethizontidae, Capromyidae, Cuniculidae, Thryonomyidae, Bathyergidae, Chinchillidae	295	0.872	0.603
out	15	Cheirogaleidae, Indriidae, Daubentoniidae, Lemuridae, Lepilemuridae	48	0.921	0.623

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.22.111344; this version posted June 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Supplemental Table 3. Predicted high-similarity bat hosts sharing with *Rhinolophus affinis* and *R. malayanus*. Species on these lists may be particularly likely to be the ultimate evolutionary origin of SARS-CoV-2, or a closely-related virus prior to recombination in an intermediate host. Predictions are made based just on the average viral sharing probability inferred for the two hosts from the phylogeography model (Trait-based 3). (* Note that the two species have high sharing probabilities with each other, potentially indicating that efforts to trace the origins of SARS-CoV-2 are already very close to their target.)

Rhinolophus affinis	Rhinolophus malayanus
1. Rhinolophus macrotis (P=0.84)	1. Rhinolophus shameli (P=0.87)
2. Rhinolophus stheno (P=0.83)	2. Rhinolophus coelophyllus (P=0.84)
3. Rhinolophus malayanus (P=0.82)	3. Rhinolophus thomasi (P=0.84)
4. Rhinolophus acuminatus (P=0.81)	4. Rhinolophus affinis (P=0.82)
5. Rhinolophus pearsonii (P=0.78)	5. Rhinolophus marshalli (P=0.82)
6. Rhinolophus shameli (P=0.78)	6. Rhinolophus pearsonii (P=0.82)
7. Rhinolophus thomasi (P=0.78)	7. Rhinolophus yunanensis (P=0.79)
8. Rhinolophus sinicus (P=0.77)	8. Rhinolophus paradoxolophus (P=0.78)
9. Rhinolophus trifoliatus (P=0.76)	9. Rhinolophus macrotis (P=0.76)
10. Rhinolophus marshalli (P=0.72)	10. Rhinolophus acuminatus (P=0.75)
11. Rhinolophus shortridgei (P=0.71)	11. Rhinolophus siamensis (P=0.75)
12. Rhinolophus luctus (P=0.7)	12. Rhinolophus rouxii (P=0.72)
13. Rhinolophus sedulus (P=0.7)	13. Rhinolophus stheno (P=0.71)
14. Rhinolophus rouxii (P=0.69)	14. Rhinolophus luctus (P=0.69)
15. Rhinolophus pusillus (P=0.68)	15. Rhinolophus trifoliatus (P=0.65)
16. Rhinolophus ferrumequinum (P=0.67)	16. Rhinolophus pusillus (P=0.62)
17. Rhinolophus lepidus (P=0.67)	17. Rhinolophus borneensis (P=0.6)
18. Hipposideros pomona (P=0.66)	18. Hipposideros lylei (P=0.59)
19. Rhinolophus celebensis (P=0.66)	19. Rhinolophus shortridgei (P=0.59)
20. Rhinolophus paradoxolophus (P=0.66)	20. Rhinolophus sinicus (P=0.59)

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.22.111344; this version posted June 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Supplemental Table 4. Predicted high-similarity non-bat hosts sharing with *Rhinolophus affinis* and *R. malayanus*. Species on these lists may be particularly suitable as stepping stones for betacoronavirus transmission from bats into humans, including potentially for SARS-CoV-2 and other SARS-like viruses. Predictions are made based just on the average viral sharing probability inferred for the two hosts from the phylogeography model (Trait-based 3). Species' binomial names are included alongside their families.

Rhinolophus affinis		***Rhinolophus malayanus***	
1. Arctonyx collaris (P=0.33)	Mustelidae	1. Arctonyx collaris (P=0.29)	Mustelidae
2. Budorcas taxicolor (P=0.33)	Bovidae	2. Herpestes urva (P=0.28)	Herpestidae
3. Viverra tangalunga (P=0.32)	Viverridae	3. Lutrogale perspicillata (P=0.28)	Mustelidae
4. Manis javanica (P=0.3)	Manidae	4. Melogale personata (P=0.27)	Mustelidae
5. Mustela altaica (P=0.3)	Mustelidae	5. Viverra megaspila (P=0.26)	Viverridae
6. Ursus thibetanus (P=0.3)	Ursidae	6. Arctictis binturong (P=0.25)	Viverridae
7. Cynogale bennettii (P=0.29)	Viverridae	7. Euroscaptor klossi (P=0.25)	Talpidae
8. Elaphodus cephalophus (P=0.29)	Cervidae	8. Lutra sumatrana (P=0.25)	Mustelidae
9. Lutrogale perspicillata (P=0.29)	Mustelidae	9. Sus scrofa (P=0.25)	Suidae
10. Viverricula indica (P=0.29)	Viverridae	10. Capricornis milneedwardsii (P=0.23)	Bovidae
11. Capricornis sumatraensis (P=0.28)	Bovidae	11. Manis javanica (P=0.23)	Manidae
12. Chimarrogale himalayica (P=0.28)	Soricidae	12. Manis pentadactyla (P=0.23)	Manidae
13. Helarctos malayanus (P=0.28)	Ursidae	13. Mustela nudipes (P=0.23)	Mustelidae
14. Herpestes javanicus (P=0.27)	Herpestidae	14. Paguma larvata (P=0.23)	Viverridae
15. Hylomys suillus (P=0.27)	Erinaceidae	15. Panthera pardus (P=0.23)	Felidae
16. Mustela kathiah (P=0.27)	Mustelidae	16. Viverra zibetha (P=0.23)	Viverridae
17. Capricornis milneedwardsii (P=0.26)	Bovidae	17. Bandicota savilei (P=0.22)	Muridae
18. Catopuma temminckii (P=0.26)	Felidae	18. Chrotogale owstoni (P=0.22)	Viverridae
19. Crocidura negligens (P=0.26)	Soricidae	19. Crocidura fuliginosa (P=0.22)	Soricidae
20. Capricornis thar (P=0.25)	Bovidae	20. Crocidura vorax (P=0.22)	Soricidae

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.22.111344; this version posted June 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Supplemental Table 5. Taxonomic scale of model training data and predictive implementation. Notes: (1) These models generated predictions of sharing with *Rhinolophus affinis* over all non-human mammals in the HP3 dataset, then subsetted to bats. (2) In these models, bat-betacoronavirus predictions are based on a subset of binary outcomes for known association with betacoronaviruses, without any other viruses included.

Model approach	Training data scale	Bat *Betacoronavirus* predictions	Mammal-wide *Betacoronavirus* predictions
Network-based 1 k-Nearest neighbors	Bat-virus	✓	
Network-based 1 k-Nearest neighbors	Mammal-virus		✓
Network-based 2 Linear filter	Bat-virus	✓	
Network-based 2 Linear filter	Mammal-virus		✓
Network-based 3 Plug-and-play	Mammal-virus[1]	✓	✓
Network-based 4 Scaled-phylogeny	Bat-virus	✓	
Network-based 4 Scaled-phylogeny	Mammal-virus		✓
Trait-based 1 Boosted regression trees	Bat-betacoronavirus[2]	✓	
Trait-based 2 Bayesian additive regression trees	Bat-betacoronavirus[2]	✓	
Trait-based 3 Neutral phylogeographic	Mammal-virus[1]	✓	✓
Hybrid 1 Two-step kernel ridge regression	Bat-betacoronavirus	✓	

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.22.111344; this version posted June 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Supplemental Table 6. Data scale of prediction, by method. Some methods use pseudoabsences to expand the scale of prediction but still only analyze existing data, with no out-of-sample inference, while others can predict freshly onto new data. (* Training data from the HP3 database uses pseudoabsences, but no new ones are generated in this study that modify the model or the bat-virus association dataset)

Model approach	Prediction on hosts without known associations (out-of-sample)	Predictive extent and use of pseudoabsences
Network-based 1 k-Nearest neighbors	No	Only predicts link probabilities among species in the association data
Network-based 2 Linear filter	No	Only predicts link probabilities among species in the association data
Network-based 3 Plug-and-play	No	Uses pseudoabsences to predict over all mammals in association data, using latent approach
Network-based 4 Scaled-phylogeny	No	Only predicts link probabilities among species in the association data
Trait-based 1 Boosted regression trees	Yes	Uses pseudoabsences for all bats in trait data to predict over all species, including those without known associations
Trait-based 2 Bayesian additive regression trees	Yes	Uses pseudoabsences for all bats in trait data to predict over all species, including those without known associations
Trait-based 3 Neutral phylogeographic	Yes	Trains on a broader network, and predicts sharing probabilities among any mammals in phylogeny and IUCN range map data
Hybrid 1 Two-step kernel ridge regression	Yes	Uses pseudoabsences for all bats in trait data to predict over all species, including those without known associations

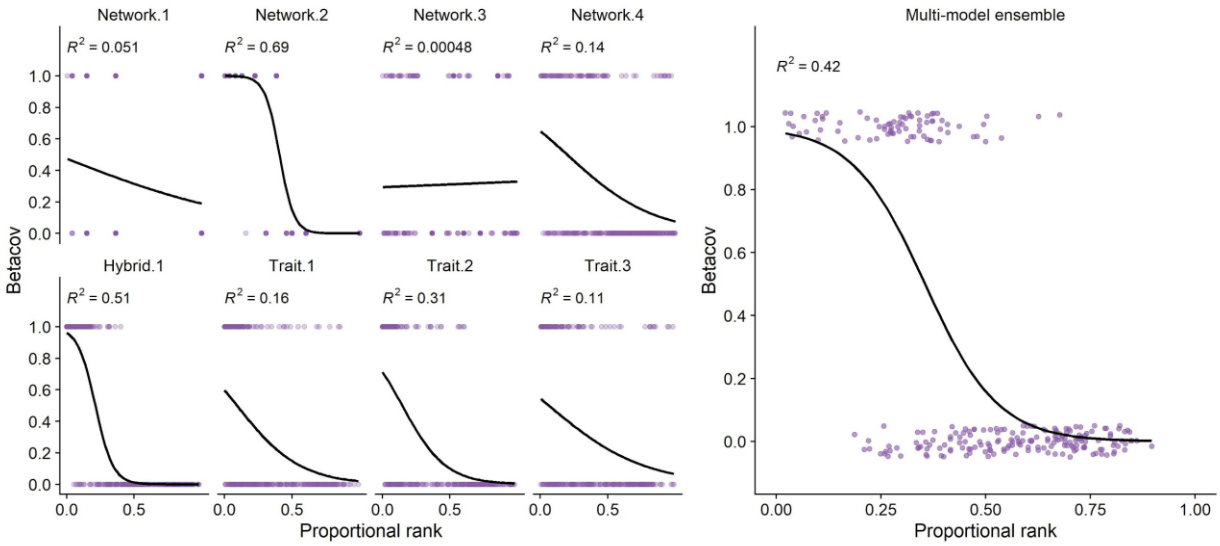
bioRxiv preprint doi: https://doi.org/10.1101/2020.05.22.111344; this version posted June 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Supplemental Table 7. Model performance for hybrid model. Area under receiver-operator curve for predicting interactions, or predicting for betacoronaviruses or for hosts, using bat traits, phylogeny or a combination of both.

	interactions	**Betacoronavirus**	**hosts (average)**
traits	0.8784	0.6742	0.7304
phylogeny	0.8848	0.73389	0.7623
both	0.8975	0.7510	0.7909

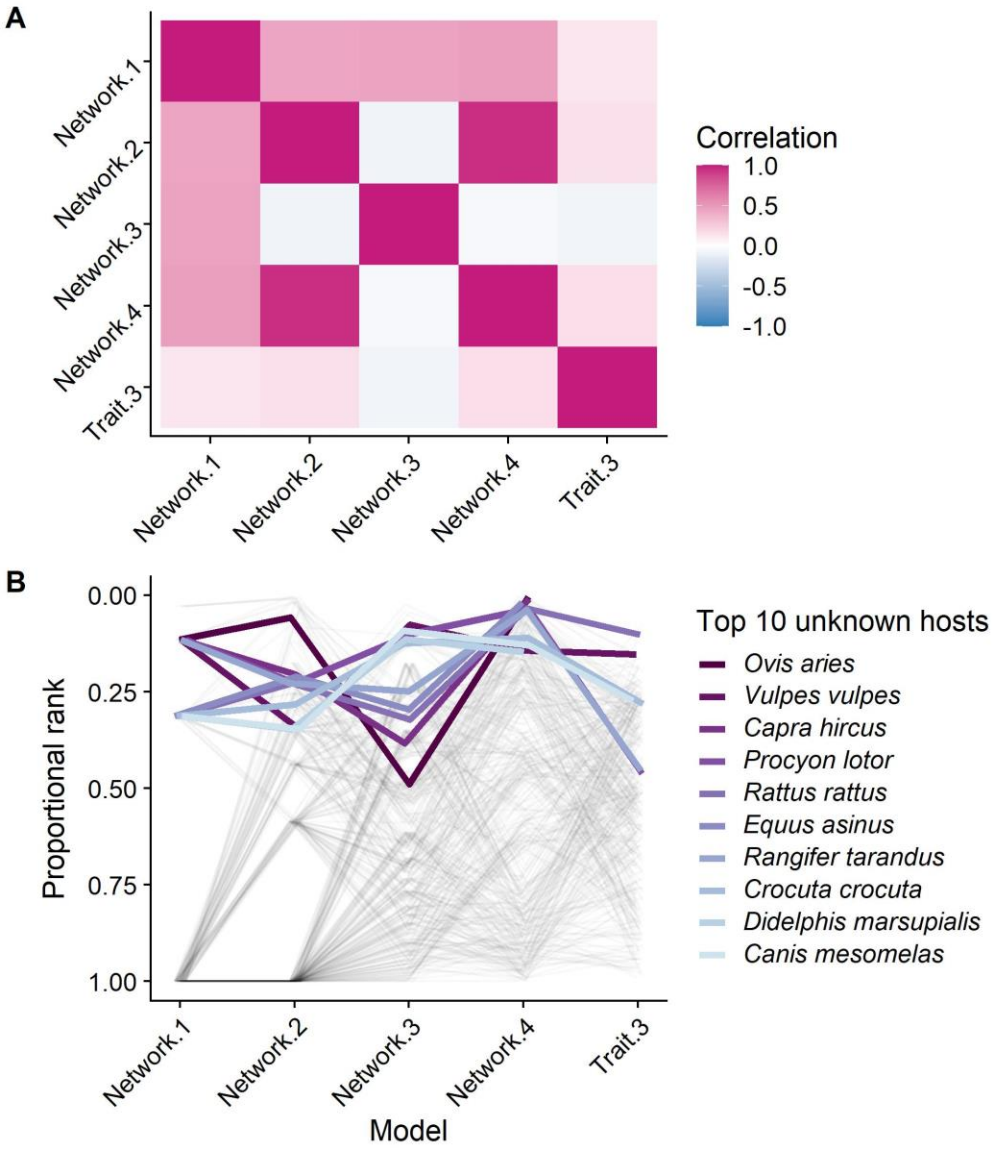
bioRxiv preprint doi: https://doi.org/10.1101/2020.05.22.111344; this version posted June 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Supplemental Figure 1. Bat models perform more strongly together than in isolation. Curves show observed betacoronavirus hosts against predicted proportional ranks from eight individual models, and incorporated into one multi-model ensemble. Black lines show a binomial GLM fit to the predicted ranks against the recorded presence or absence of known betacoronavirus associations. Points are jittered to reduce overlap.



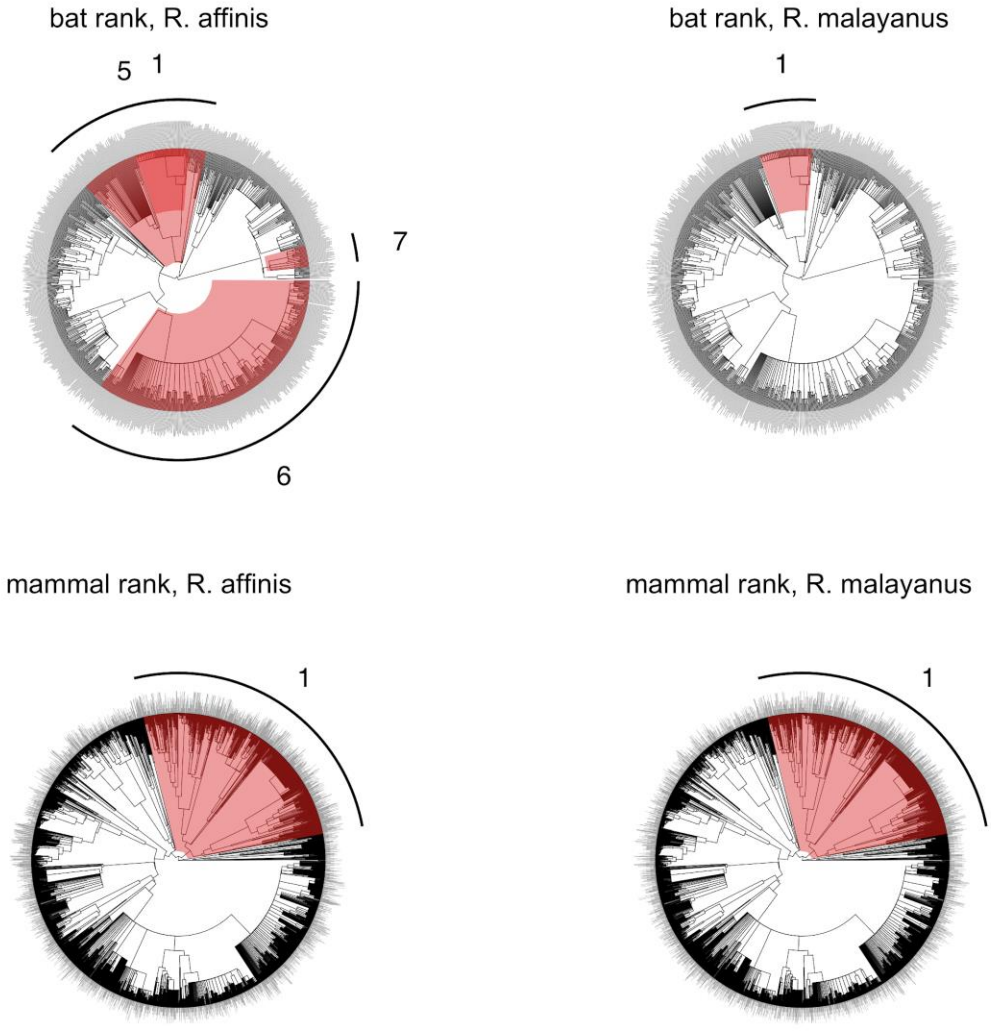
bioRxiv preprint doi: https://doi.org/10.1101/2020.05.22.111344; this version posted June 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Supplemental Figure 2. Poor concordance among predictive models for mammal hosts of betacoronaviruses. The pairwise Spearman's rank correlations between models' ranked species-level predictions were generally low (A). In-sample predictions varied significantly and heavily prioritized domestic animals and well-studied hosts (B). The ten species with the highest mean proportional ranks across all models are highlighted in shades of purple. Only in-sample predictions are displayed because only one model (Trait-based 3) was able to predict out of sample for all mammals.



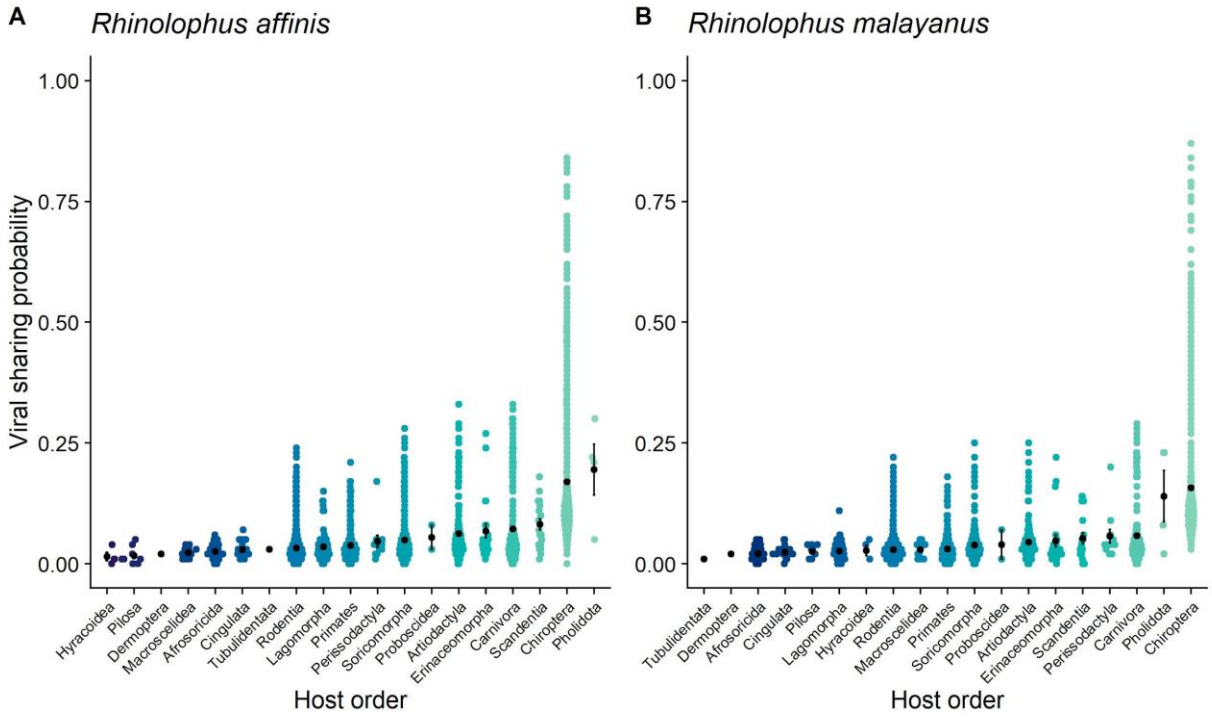
bioRxiv preprint doi: https://doi.org/10.1101/2020.05.22.111344; this version posted June 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Supplemental Figure 3. Results of phylogenetic factorization applied to predicted ranks of virus sharing with *Rhinolophus affinis* and *Rhinolophus malayanus*. Colored regions indicate clades identified as significantly different in their predicted rank compared to the paraphyletic remainder; those more likely to share a virus with the *Rhinolophus* are shown in red, whereas those less likely to share a virus are shown in blue. Bar height indicates predicted rank (higher values = lower rank score, more likely share virus). Results are displayed for bats and remaining mammals separately. Mammal-wide clades with high propensities to share viruses with *R. affinis* based solely on their phylogeography included the treeshrews (Scandentia), Old World monkeys (Colobinae), and both tufted and barking deer (Muntiacini).



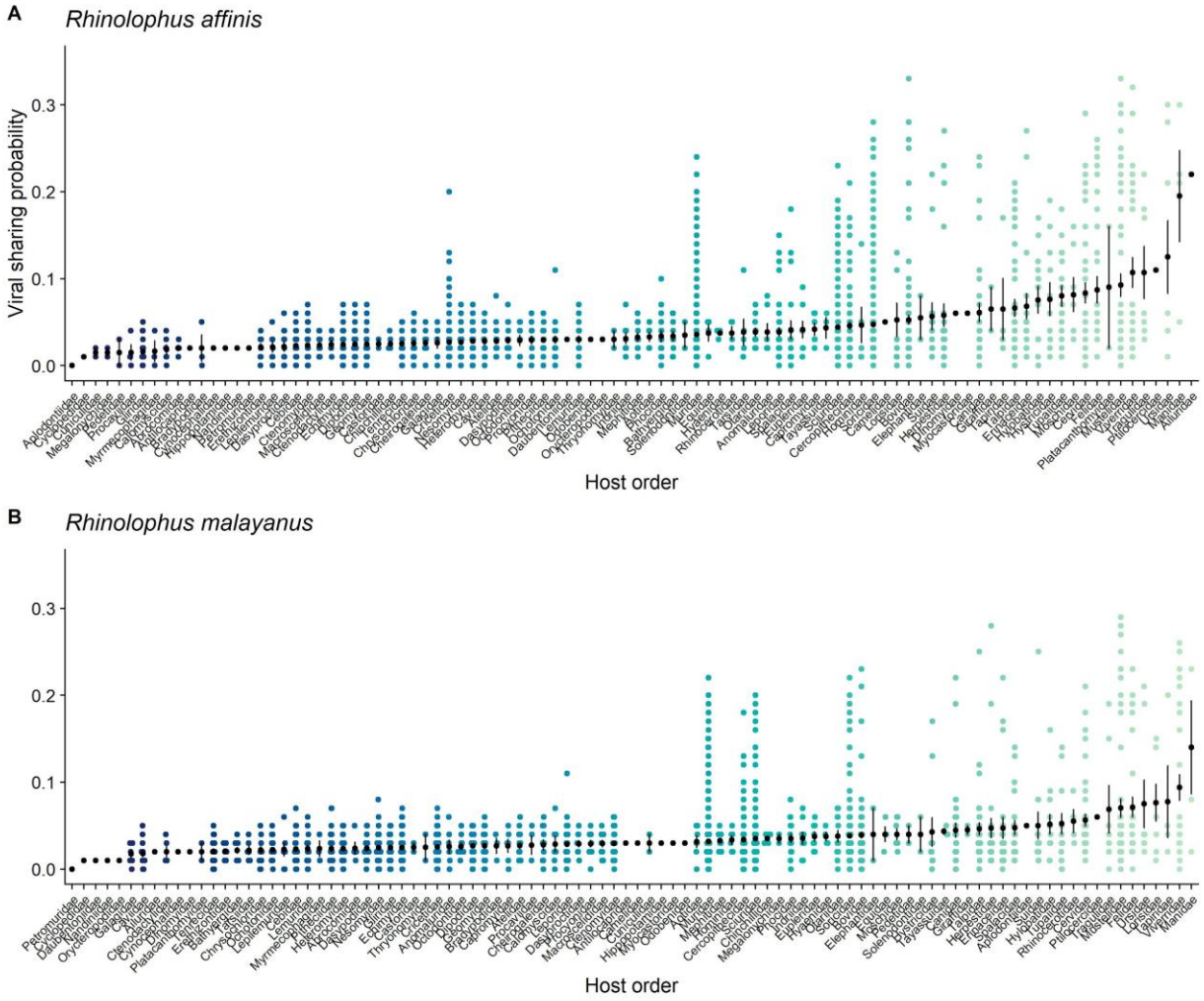
bioRxiv preprint doi: https://doi.org/10.1101/2020.05.22.111344; this version posted June 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Supplemental Figure 4. Predicted species-level sharing probabilities of A) *Rhinolophus affinis* and B) *Rhinolophus malayanus*, calculated according to the phylogeographic viral sharing model[48]. Each coloured point is a mammal species. Black points and error bars denote means and standard errors for each order. Mammal orders are arranged according to their mean sharing probability.



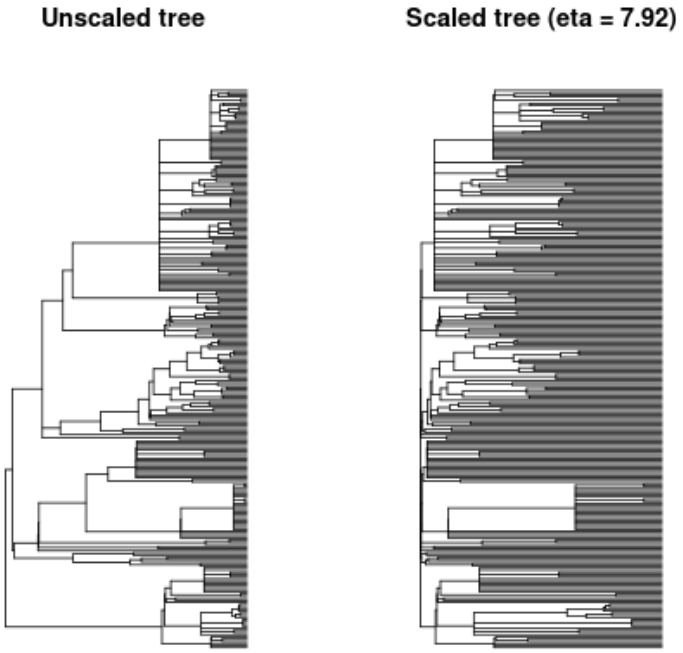
bioRxiv preprint doi: https://doi.org/10.1101/2020.05.22.111344; this version posted June 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Supplemental Figure 5. Predicted species-level sharing probabilities of A) *Rhinolophus affinis* and B) *Rhinolophus malayanus*, calculated according to the phylogeographic viral sharing model^. Each coloured point is a mammal species. Black points and error bars denote means and standard errors for each order. Mammal orders are arranged according to their mean sharing probability.



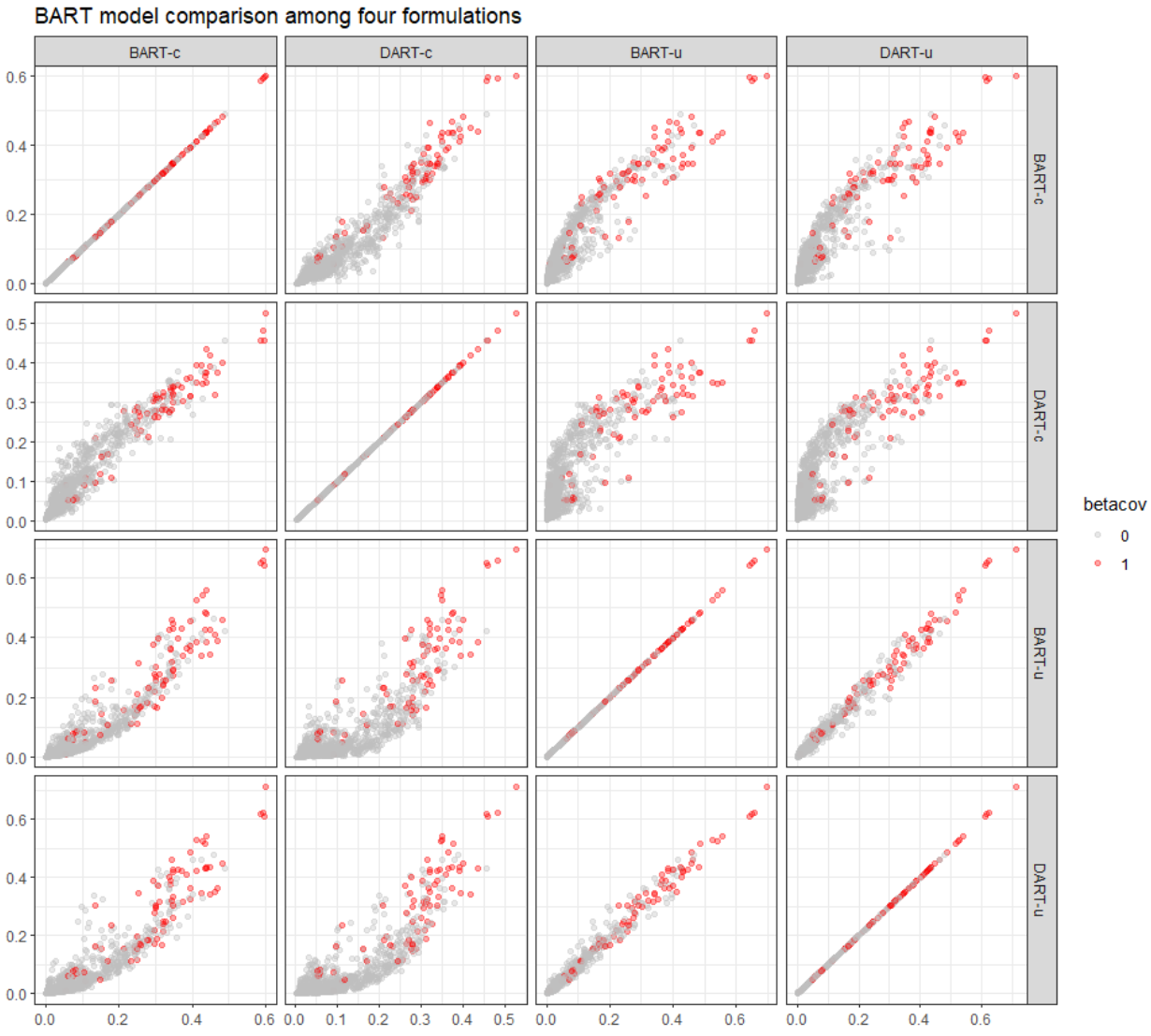
bioRxiv preprint doi: https://doi.org/10.1101/2020.05.22.111344; this version posted June 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Supplemental Figure 6. To account for uncertainty in the phylogenetic distances among hosts, the scaled-phylogeny model estimates a tree scaling parameter (eta) based on an early-burst model of evolution. On the left is the unscaled bat phylogeny for the hosts in the bat-virus genera network, and on the right is the same tree rescaled according to mean estimated scaling parameter (eta = 7.92). Eta values above 1 indicate accelerating evolution, suggesting less phylogenetic conservatism in host-virus associations among closely related taxa than would be predicted by a Brownian-motion model on the unscaled tree.



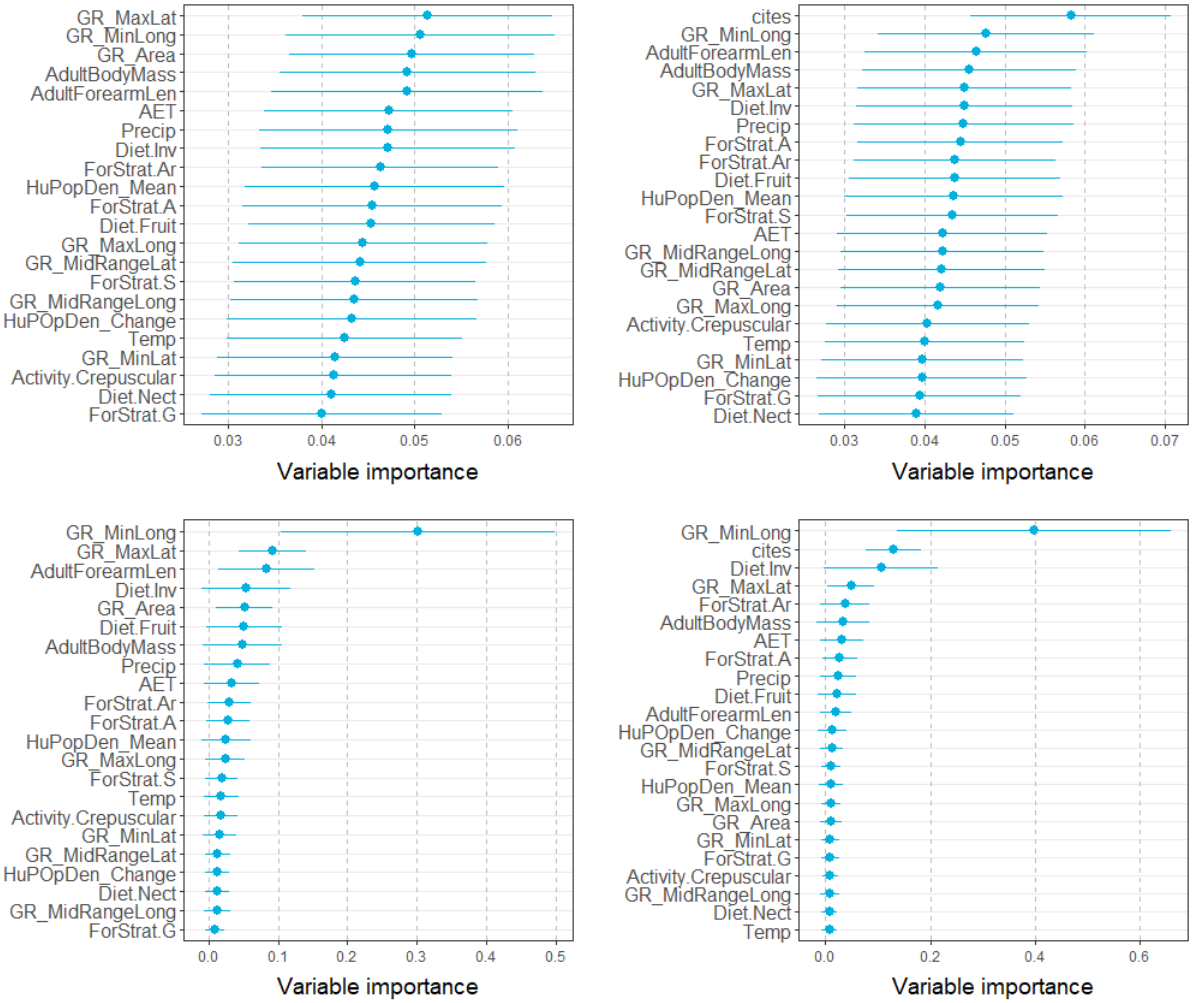
bioRxiv preprint doi: https://doi.org/10.1101/2020.05.22.111344; this version posted June 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Supplemental Figure 7. Four formulations of Bayesian additive regression tree (BART) models produce slightly different results, but largely agree. Two models use baseline BART, while two models use a Dirichlet prior on variable importance (DART). Two are uncorrected for sampling bias (u) while two are corrected using citation counts (c). In the final main-text model ensemble, we present a DART model including correction for citation bias, which penalizes overfitting and spurious patterns two ways and leads to predictions with a lower total correlation with the data, but a still-high performance (AUC = 0.90).



bioRxiv preprint doi: https://doi.org/10.1101/2020.05.22.111344; this version posted June 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

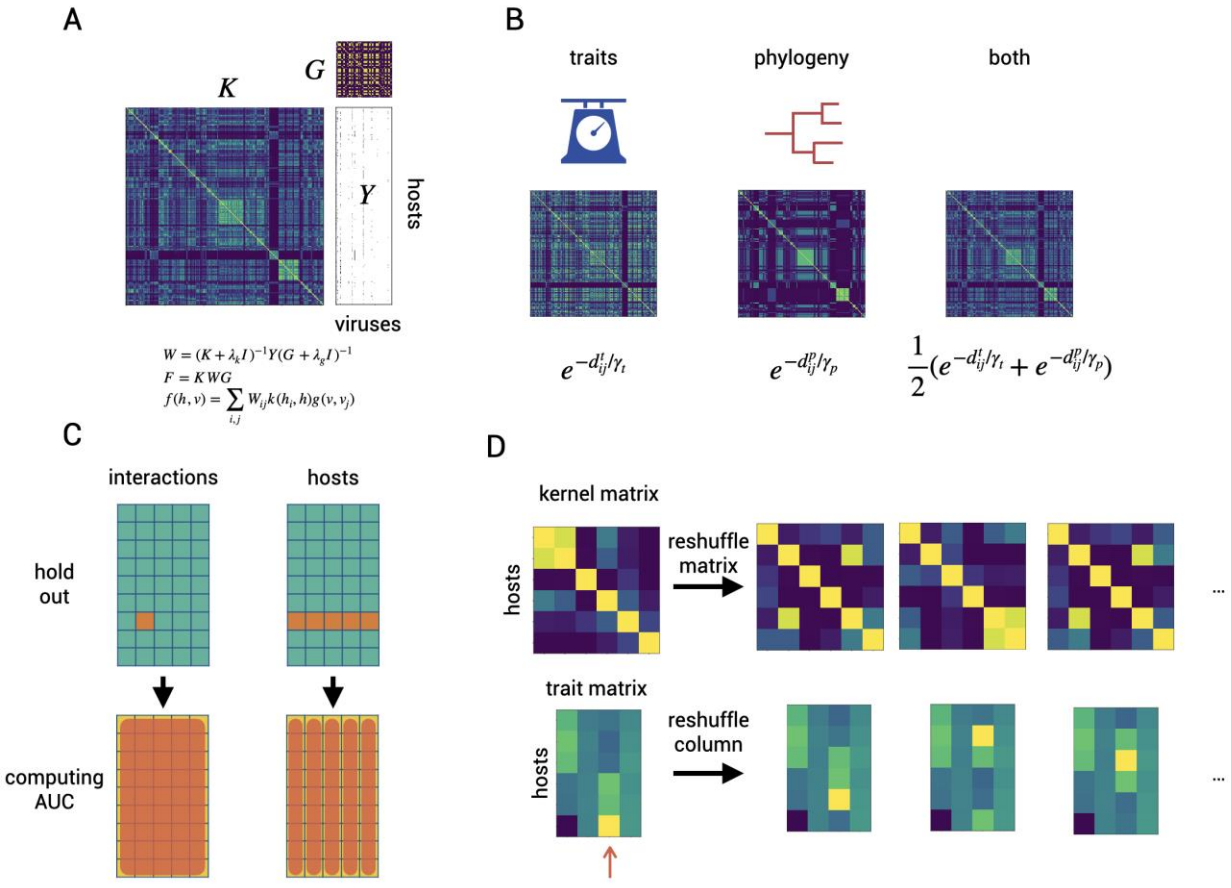
1119 **Supplemental Figure 8.** Variable importance plots for the Bayesian additive regression tree
1120 models with uniform variable importance prior (top) versus Dirichlet prior (bottom), without (left)
1121 and with (right) correction for citations.



1122

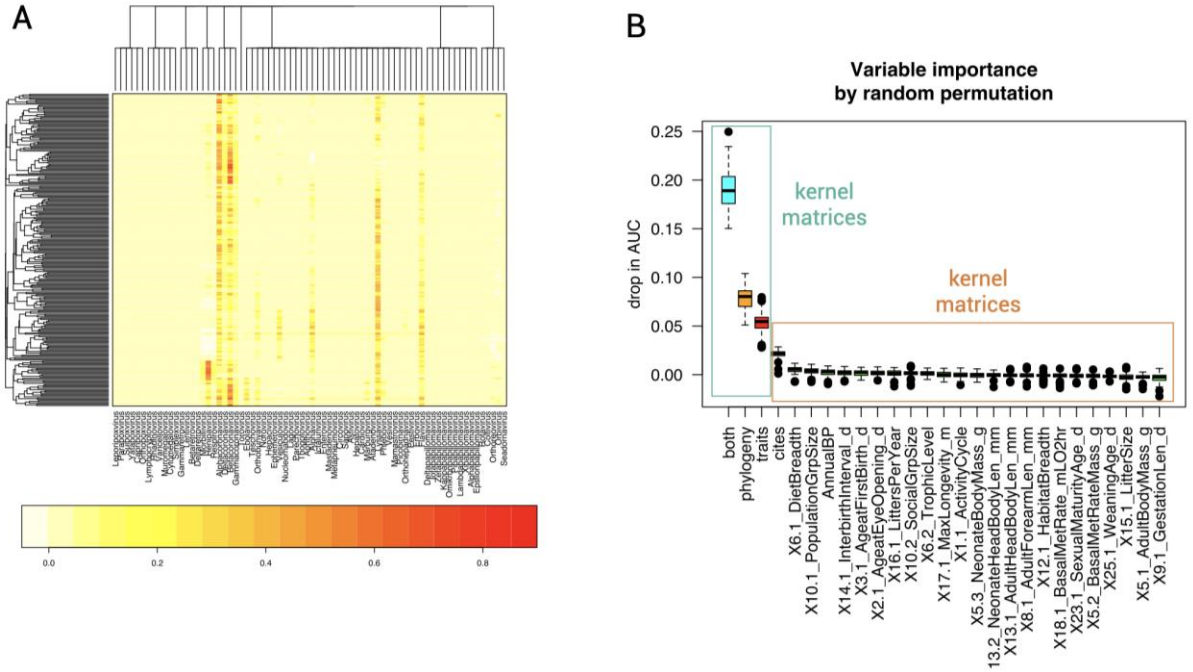
bioRxiv preprint doi: https://doi.org/10.1101/2020.05.22.111344; this version posted June 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Supplemental Figure 9. Overview of the Two-Step Kernel Ridge Regression method. A. The incidence matrix *Y* is modelled by two kernel matrices *K* and *G*, describing the hosts (here: bats) and viruses, respectively. The equations for fitting the model and making predictions are given below. B. Computing the host kernel matrix is done based on traits, phylogeny or both. A standard radial basis kernel is used to transform a distance to a valid kernel. The two kernels are combined by averaging. C. The models are validated by specialized cross-validation. The setting interactions leaves one interaction out at a time and repeats this for every interaction. AUC is computed in micro-fashion. In the setting hosts, we leave out one host at a time and compute the AUC per virus, averaging the individual results. D. To assess the variable importance, we first randomly reshuffle the kernel matrices and monitor how the performance of a model fitted on these data drops. In a more fine-grained approach, we just reshuffle a single trait (indicated by the red arrow) and compare the resulting performances.



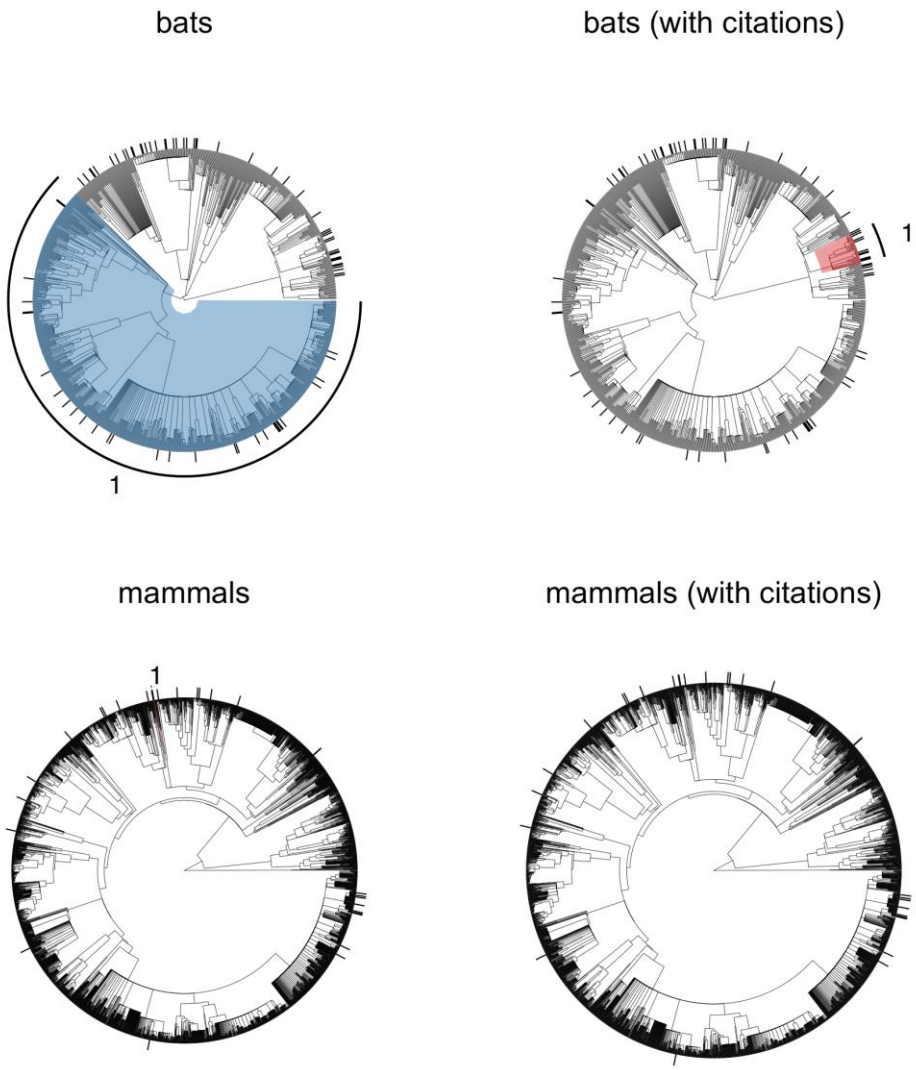
bioRxiv preprint doi: https://doi.org/10.1101/2020.05.22.111344; this version posted June 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

1137 **Supplemental Figure 10. Illustration of the TSKRR model and the variable importance scores.** A.
1138 Heatmap of the prediction scores, together with a hierarchical clustering based on the kernel
1139 distance of *K* and *G*. B. Variable importance computed by random permutations of the kernel
1140 matrices or a single trait.



bioRxiv preprint doi: https://doi.org/10.1101/2020.05.22.111344; this version posted June 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Supplemental Figure 11. Results of phylogenetic factorization applied to binomial betacoronavirus data across bats (top) and other mammals (bottom), using raw data (left) and after weighting by citation counts (right). Any significant clades (5% family-wise error rate) are displayed in colored sharing on the phylogeny. Bars indicate betacoronavirus detection, and clades are colored by having more (red) or fewer (blue) positive species.



bioRxiv preprint doi: https://doi.org/10.1101/2020.05.22.111344; this version posted June 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Bibliography

1. Anthony, S. J. *et al.* Global patterns in coronavirus diversity. *Virus Evol* **3**, vex012 (2017).
2. Denison, M. R., Graham, R. L., Donaldson, E. F., Eckerle, L. D. & Baric, R. S. Coronaviruses: an RNA proofreading machine regulates replication fidelity and diversity. *RNA Biol.* **8**, 270–279 (2011).
3. Ren, W. *et al.* Full-length genome sequences of two SARS-like coronaviruses in horseshoe bats and genetic variation analysis. *J. Gen. Virol.* **87**, 3355–3359 (2006).
4. Li, W. *et al.* Bats are natural reservoirs of SARS-like coronaviruses. *Science* **310**, 676–679 (2005).
5. Yang, X.-L. *et al.* Isolation and Characterization of a Novel Bat Coronavirus Closely Related to the Direct Progenitor of Severe Acute Respiratory Syndrome Coronavirus. *J. Virol.* **90**, 3253–3256 (2015).
6. Zhou, P. *et al.* A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature* **579**, 270–273 (2020).
7. Guan, Y. *et al.* Isolation and characterization of viruses related to the SARS coronavirus from animals in southern China. *Science* **302**, 276–278 (2003).
8. Hu, B. *et al.* Discovery of a rich gene pool of bat SARS-related coronaviruses provides new insights into the origin of SARS coronavirus. *PLoS Pathog.* **13**, e1006698 (2017).
9. Memish, Z. A. *et al.* Middle East respiratory syndrome coronavirus in bats, Saudi Arabia. *Emerg. Infect. Dis.* **19**, 1819–1823 (2013).
10. Wang, Q. *et al.* Bat origins of MERS-CoV supported by bat coronavirus HKU4 usage of human receptor CD26. *Cell Host Microbe* **16**, 328–337 (2014).
11. Yang, Y. *et al.* Receptor usage and cell entry of bat coronavirus HKU4 provide insight into bat-to-human transmission of MERS coronavirus. *Proc. Natl. Acad. Sci. U. S. A.* **111**, 12516–12521 (2014).
12. Hu, B., Ge, X., Wang, L.-F. & Shi, Z. Bat origin of human coronaviruses. *Virology Journal* vol. 12 (2015).
13. Anthony, S. J. *et al.* Further Evidence for Bats as the Evolutionary Source of Middle East Respiratory Syndrome Coronavirus. *MBio* **8**, (2017).
14. Anthony, S. J. *et al.* Coronaviruses in bats from Mexico. *J. Gen. Virol.* **94**, 1028–1038 (2013).
15. Yang, L. *et al.* MERS-related betacoronavirus in Vespertilio superans bats, China. *Emerg. Infect. Dis.* **20**, 1260–1262 (2014).
16. Zhou, H. *et al.* A novel bat coronavirus reveals natural insertions at the S1/S2 cleavage site of the Spike protein and a possible recombinant origin of HCoV-19. doi:10.1101/2020.03.02.974139.
17. Nielsen, R., Wang, H. & Pipes, L. Synonymous mutations and the molecular evolution of SARS-Cov-2 origins. doi:10.1101/2020.04.20.052019.
18. Lam, T. T.-Y. *et al.* Identifying SARS-CoV-2 related coronaviruses in Malayan pangolins. *Nature* (2020) doi:10.1038/s41586-020-2169-0.
19. Xiao, K. *et al.* Isolation of SARS-CoV-2-related coronavirus from Malayan pangolins. *Nature* (2020) doi:10.1038/s41586-020-2313-x.

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.22.111344; this version posted June 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

20. Zhang, T., Wu, Q. & Zhang, Z. Probable Pangolin Origin of SARS-CoV-2 Associated with the COVID-19 Outbreak. *Curr. Biol.* **30**, 1578 (2020).
21. Andersen, K. G., Rambaut, A., Ian Lipkin, W., Holmes, E. C. & Garry, R. F. The proximal origin of SARS-CoV-2. *Nature Medicine* vol. 26 450–452 (2020).
22. Viana, M. *et al.* Assembling evidence for identifying reservoirs of infection. *Trends Ecol. Evol.* **29**, 270–279 (2014).
23. Plowright, R. K. *et al.* Prioritizing surveillance of Nipah virus in India. *PLoS Negl. Trop. Dis.* **13**, e0007393 (2019).
24. 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).
25. Becker, D. J., Washburne, A. D., Faust, C. L., Mordecai, E. A. & Plowright, R. K. The problem of scale in the prediction and management of pathogen spillover. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **374**, 20190224 (2019).
26. Becker, D. J. & Han, B. A. The macroecology and evolution of avian competence for Borrelia burgdorferi. doi:10.1101/2020.04.15.040352.
27. Han, B. A. *et al.* Undiscovered Bat Hosts of Filoviruses. *PLoS Negl. Trop. Dis.* **10**, e0004815 (2016).
28. Han, B. A. *et al.* Confronting data sparsity to identify potential sources of Zika virus spillover infection among primates. *Epidemics* **27**, 59–65 (2019).
29. Washburne, A. D. *et al.* Taxonomic patterns in the zoonotic potential of mammalian viruses. *PeerJ* **6**, e5979 (2018).
30. Olival, K. J. *et al.* Host and viral traits predict zoonotic spillover from mammals. *Nature* **546**, 646–650 (2017).
31. Fritz, S. A., Bininda-Emonds, O. R. P. & Purvis, A. Geographical variation in predictors of mammalian extinction risk: big is bad, but only in the tropics. *Ecol. Lett.* **12**, 538–549 (2009).
32. Jones, K. E. *et al.* PanTHERIA: a species-level database of life history, ecology, and geography of extant and recently extinct mammals. *Ecology* vol. 90 2648–2648 (2009).
33. Wilman, H. *et al.* EltonTraits 1.0: Species-level foraging attributes of the world's birds and mammals. *Ecology* vol. 95 2027–2027 (2014).
34. Trifonova, N. *et al.* Spatio-temporal Bayesian network models with latent variables for revealing trophic dynamics and functional networks in fisheries ecology. *Ecological Informatics* vol. 30 142–158 (2015).
35. Rohr, R. P., Scherer, H., Kehrli, P., Mazza, C. & Bersier, L.-F. Modeling food webs: exploring unexplained structure using latent traits. *Am. Nat.* **176**, 170–177 (2010).
36. Dallas, T., Park, A. W. & Drake, J. M. Predicting cryptic links in host-parasite networks. *PLoS Comput. Biol.* **13**, e1005557 (2017).
37. Han, B. A., Schmidt, J. P., Bowden, S. E. & Drake, J. M. Rodent reservoirs of future zoonotic diseases. *Proceedings of the National Academy of Sciences* vol. 112 7039–7044 (2015).
38. Brandão, P. E. *et al.* A coronavirus detected in the vampire bat Desmodus rotundus. *Braz. J. Infect. Dis.* **12**, 466–468 (2008).

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.22.111344; this version posted June 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

39. Corman, V. M. *et al.* Highly diversified coronaviruses in neotropical bats. *J. Gen. Virol.* **94**, 1984–1994 (2013).
40. Moreira-Soto, A. *et al.* Neotropical bats from Costa Rica harbour diverse coronaviruses. *Zoonoses Public Health* **62**, 501–505 (2015).
41. Wang, L. *et al.* Discovery and genetic analysis of novel coronaviruses in least horseshoe bats in southwestern China. *Emerg. Microbes Infect.* **6**, e14 (2017).
42. Lin, X.-D. *et al.* Extensive diversity of coronaviruses in bats from China. *Virology* vol. 507 1–10 (2017).
43. Wacharapluesadee, S. *et al.* Diversity of coronavirus in bats from Eastern Thailand. *Virol. J.* **12**, 57 (2015).
44. Guy, C., Ratcliffe, J. M. & Mideo, N. The influence of bat ecology on viral diversity and reservoir status. *Ecol. Evol.* **2008**, 209 (2020).
45. Washburne, A. D. *et al.* Phylofactorization: a graph partitioning algorithm to identify phylogenetic scales of ecological data. *Ecol. Monogr.* **89**, e01353 (2019).
46. Crowley, D., Becker, D., Washburne, A. & Plowright, R. Identifying Suspect Bat Reservoirs of Emerging Infections. *Vaccines* vol. 8 228 (2020).
47. Damas, J., Hughes, G. M., Keough, K. C. & Painter, C. A. Broad Host Range of SARS-CoV-2 Predicted by Comparative and Structural Analysis of ACE2 in Vertebrates. *bioRxiv* (2020).
48. 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).
49. Wang, M. *et al.* SARS-CoV Infection in a Restaurant from Palm Civet. *Emerging Infectious Diseases* vol. 11 1860–1865 (2005).
50. Song, H.-D. *et al.* Cross-host evolution of severe acute respiratory syndrome coronavirus in palm civet and human. *Proc. Natl. Acad. Sci. U. S. A.* **102**, 2430–2435 (2005).
51. Dougherty, E. R., Seidel, D. P., Carlson, C. J., Spiegel, O. & Getz, W. M. Going through the motions: incorporating movement analyses into disease research. *Ecol. Lett.* **21**, 588–604 (2018).
52. Lehmann, D. *et al.* Pangolins and bats living together in underground burrows in Lopé National Park, Gabon. *African Journal of Ecology* (2020) doi:10.1111/aje.12759.
53. Oreshkova, N. *et al.* SARS-CoV2 infection in farmed mink, Netherlands, April 2020. doi:10.1101/2020.05.18.101493.
54. Kim, Y.-I. *et al.* Infection and Rapid Transmission of SARS-CoV-2 in Ferrets. *Cell Host Microbe* **27**, 704–709.e2 (2020).
55. Shi, J. *et al.* Susceptibility of ferrets, cats, dogs, and other domesticated animals to SARS–coronavirus 2. *Science* (2020) doi:10.1126/science.abb7015.
56. Yang, X.-L. *et al.* Genetically Diverse Filoviruses in Rousettus and Eonycteris spp. Bats, China, 2009 and 2015. *Emerg. Infect. Dis.* **23**, 482–486 (2017).
57. Seifert, S. N. *et al.* Rousettus aegyptiacus Bats Do Not Support Productive Nipah Virus Replication. *The Journal of Infectious Diseases* vol. 221 S407–S413 (2020).
58. Wacharapluesadee, S. *et al.* Longitudinal study of age-specific pattern of coronavirus infection in Lyle's flying fox (Pteropus lylei) in Thailand. *Virol. J.* **15**, 38 (2018).
59. Latinne, A. *et al.* Origin and cross-species transmission of bat coronaviruses in China.

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.22.111344; this version posted June 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

doi:10.1101/2020.05.31.116061.

60. Yang, L. *et al.* Novel SARS-like betacoronaviruses in bats, China, 2011. *Emerg. Infect. Dis.* **19**, 989–991 (2013).
61. Geldenhuys, M. *et al.* A metagenomic viral discovery approach identifies potential zoonotic and novel mammalian viruses in Neoromicia bats within South Africa. *PLoS One* **13**, e0194527 (2018).
62. Memish, Z. A. *et al.* Middle East respiratory syndrome coronavirus in bats, Saudi Arabia. *Emerg. Infect. Dis.* **19**, 1819–1823 (2013).
63. Luo, Y. *et al.* Longitudinal Surveillance of Betacoronaviruses in Fruit Bats in Yunnan Province, China During 2009-2016. *Virol. Sin.* **33**, 87–95 (2018).
64. Peel, A. J. *et al.* Synchronous shedding of multiple bat paramyxoviruses coincides with peak periods of Hendra virus spillover. *Emerg. Microbes Infect.* **8**, 1314–1323 (2019).
65. de Souza Cortez J. L. Dunnum A. W. Ferguson F. A. Anwarali Khan D. L. Paul D. M. Reeder N. B. Simmons B. M. Thiers C. W. Thompson N S. Upham M. P. M. Vanhove P. W. Webala M. Weksler R. Yanagihara P. S. Soltis., C. J. A. S. A. B. A. J. B. C. A. C. B. M. B. Integrating biodiversity infrastructure into pathogen discovery and mitigation of epidemic infectious diseases. *Bioscience* (2020) doi:biaa064.
66. Kingston, T. *et al.* Networking networks for global bat conservation. in *Bats in the Anthropocene: Conservation of Bats in a Changing World* 539–569 (Springer, Cham, 2016).
67. Phelps, K. L. *et al.* Bat Research Networks and Viral Surveillance: Gaps and Opportunities in Western Asia. *Viruses* **11**, (2019).
68. Teeling, E. C. *et al.* Bat Biology, Genomes, and the Bat1K Project: To Generate Chromosome-Level Genomes for All Living Bat Species. *Annu Rev Anim Biosci* **6**, 23–46 (2018).
69. Mandl, J. N., Schneider, C., Schneider, D. S. & Baker, M. L. Going to Bat(s) for Studies of Disease Tolerance. *Front. Immunol.* **9**, 2112 (2018).
70. Skirmuntt, E. C., Escalera-Zamudio, M., Teeling, E. C., Smith, A. & Katzourakis, A. The Potential Role of Endogenous Viral Elements in the Evolution of Bats as Reservoirs for Zoonotic Viruses. *Annu Rev Virol* (2020) doi:10.1146/annurev-virology-092818-015613.
71. Jebb, D. *et al.* Six new reference-quality bat genomes illuminate the molecular basis and evolution of bat adaptations. (2019).
72. Gervasi, S. S., Civitello, D. J., Kilvitis, H. J. & Martin, L. B. The context of host competence: a role for plasticity in host–parasite dynamics. *Trends Parasitol.* **31**, 419–425 (2015).
73. Martin, L. B., Burgan, S. C., Adelman, J. S. & Gervasi, S. S. Host Competence: An Organismal Trait to Integrate Immunology and Epidemiology. *Integr. Comp. Biol.* **56**, 1225–1237 (2016).
74. Callaway, E. & Cyranoski, D. Why snakes probably aren't spreading the new China virus. *Nature* (2020) doi:10.1038/d41586-020-00180-8.
75. Gong, Y., Wen, G., Jiang, J. & Feng, X. Complete title: Codon bias analysis may be insufficient for identifying host(s) of a novel virus. *J. Med. Virol.* (2020) doi:10.1002/jmv.25977.
76. Zhao, H. COVID-19 drives new threat to bats in China. *Science* **367**, 1436 (2020).
77. Fenton, M. B. *et al.* Knowledge gaps about rabies transmission from vampire bats to

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.22.111344; this version posted June 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

humans. *Nature Ecology & Evolution* **4**, 517–518 (2020).
78. López-Baucells, A., Rocha, R. & Fernández-Llamazares, Á. When bats go viral: negative framings in virological research imperil bat conservation. *Mamm. Rev.* **48**, 62–66 (2018).
79. O'Shea, T. J., Cryan, P. M., Hayman, D. T. S., Plowright, R. K. & Streicker, D. G. Multiple mortality events in bats: a global review. *Mamm. Rev.* **46**, 175–190 (2016).
80. MB Fenton, S Mubareka, SM Tsang, NB Simmons, DJ Becker. COVID-19 and threats to bats. *FACETS* in press (2020).
81. Aguiar, L. M. S., Brito, D. & Machado, R. B. Do current vampire bat (Desmodus rotundus) population control practices pose a threat to Dekeyser's nectar bat's (Lonchophylla dekeyseri) long-term persistence in the Cerrado? *Acta Chiropt.* **12**, 275–282 (2010).
82. Streicker, D. G. *et al.* Ecological and anthropogenic drivers of rabies exposure in vampire bats: implications for transmission and control. *Proc. Biol. Sci.* **279**, 3384–3392 (2012).
83. Blackwood, J. C., Streicker, D. G., Altizer, S. & Rohani, P. Resolving the roles of immunity, pathogenesis, and immigration for rabies persistence in vampire bats. *Proc. Natl. Acad. Sci. U. S. A.* **110**, 20837–20842 (2013).
84. Frick, W. F. *et al.* An emerging disease causes regional population collapse of a common North American bat species. *Science* **329**, 679–682 (2010).
85. Sabir, J. S. M. *et al.* Co-circulation of three camel coronavirus species and recombination of MERS-CoVs in Saudi Arabia. *Science* **351**, 81–84 (2016).
86. Guth, S., Visher, E., Boots, M. & Brook, C. E. Host phylogenetic distance drives trends in virus virulence and transmissibility across the animal–human interface. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **374**, 20190296 (2019).
87. Redondo, R. A. F., Brina, L. P. S., Silva, R. F., Ditchfield, A. D. & Santos, F. R. Molecular systematics of the genus Artibeus (Chiroptera: Phyllostomidae). *Mol. Phylogenet. Evol.* **49**, 44–58 (2008).
88. Bouchard, S. Chaerephon pumilus. *Mammalian Species* 1–6 (1998).
89. Hoofer, S. R., Van Den Bussche, R. A. & HoráČek, I. Generic Status of the American Pipistrelles (Vespertilionidae) with Description of a New Genus. *J. Mammal.* **87**, 981–992 (2006).
90. Desjardins-Proulx, P., Laigle, I., Poisot, T. & Gravel, D. Ecological interactions and the Netflix problem. *PeerJ* **5**, e3644 (2017).
91. Stock, M., Poisot, T., Waegeman, W. & De Baets, B. Linear filtering reveals false negatives in species interaction data. *Sci. Rep.* **7**, 45908 (2017).
92. Drake, J. M. & Richards, R. L. Estimating environmental suitability. *Ecosphere* vol. 9 e02373 (2018).
93. Dallas, T. A., Carlson, C. J. & Poisot, T. Testing predictability of disease outbreaks with a simple model of pathogen biogeography. *R Soc Open Sci* **6**, 190883 (2019).
94. Elmasri, M., Farrell, M. J., Jonathan Davies, T. & Stephens, D. A. A hierarchical Bayesian model for predicting ecological interactions using scaled evolutionary relationships. *The Annals of Applied Statistics* vol. 14 221–240 (2020).
95. Cadotte, M. W. *et al.* Phylogenetic diversity metrics for ecological communities: integrating species richness, abundance and evolutionary history. *Ecol. Lett.* **13**, 96–105 (2010).

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.22.111344; this version posted June 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

96. Park, A. W. *et al.* Characterizing the phylogenetic specialism–generalism spectrum of mammal parasites. *Proceedings of the Royal Society B: Biological Sciences* vol. 285 20172613 (2018).
97. Harvey, P. H. & Pagel, M. D. *The comparative method in evolutionary biology*. (Oxford University Press, USA, 1998).
98. Harmon, L. J. *et al.* Early bursts of body size and shape evolution are rare in comparative data. *Evolution* **64**, 2385–2396 (2010).
99. Mollentze, N. & Streicker, D. G. Viral zoonotic risk is homogenous among taxonomic orders of mammalian and avian reservoir hosts. *Proc. Natl. Acad. Sci. U. S. A.* **117**, 9423–9430 (2020).
100. Schmidt, J. P. *et al.* Ecological indicators of mammal exposure to Ebolavirus. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **374**, 20180337 (2019).
101. Pandit, P. S. *et al.* Predicting wildlife reservoirs and global vulnerability to zoonotic Flaviviruses. *Nat. Commun.* **9**, 5425 (2018).
102. Evans, M. V., Dallas, T. A., Han, B. A., Murdock, C. C. & Drake, J. M. Data-driven identification of potential Zika virus vectors. *Elife* **6**, (2017).
103. Yang, L. H. & Han, B. A. Data-driven predictions and novel hypotheses about zoonotic tick vectors from the genus Ixodes. *BMC Ecol.* **18**, 7 (2018).
104. Elith, J., Leathwick, J. R. & Hastie, T. A working guide to boosted regression trees. *J. Anim. Ecol.* **77**, 802–813 (2008).
105. Carlson, C. J. embarcadero: Species distribution modelling with Bayesian additive regression trees in R. doi:10.1101/774604.
106. Chipman, H. A., George, E. I. & McCulloch, R. E. BART: Bayesian additive regression trees. *The Annals of Applied Statistics* vol. 4 266–298 (2010).
107. Chen, W. *et al.* The illegal exploitation of hog badgers (Arctonyx collaris) in China: genetic evidence exposes regional population impacts. *Conservation Genetics Resources* vol. 7 697–704 (2015).
108. Stock, M., Pahikkala, T., Airola, A., De Baets, B. & Waegeman, W. A Comparative Study of Pairwise Learning Methods Based on Kernel Ridge Regression. *Neural Comput.* **30**, 2245–2283 (2018).
109. Stock, M., Pahikkala, T., Airola, A., Waegeman, W. & De Baets, B. Algebraic shortcuts for leave-one-out cross-validation in supervised network inference. *Brief. Bioinform.* (2018) doi:10.1093/bib/bby095.
110. Stekhoven, D. J. & Bühlmann, P. MissForest—non-parametric missing value imputation for mixed-type data. *Bioinformatics* **28**, 112–118 (2012).
111. Schölkopf, B., Smola, A. J., Managing Director of the Max Planck Institute for Biological Cybernetics in Tubingen Germany Profe Bernhard Scholkopf & Bach, F. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. (MIT Press, 2002).
112. Foley, N. M., Goodman, S. M., Whelan, C. V., Puechmaille, S. J. & Teeling, E. Towards navigating the Minotaur's labyrinth: cryptic diversity and taxonomic revision within the speciose genus Hipposideros (Hipposideridae). *Acta Chiropt.* **19**, 1–18 (2017).