Electronic supplemental material for

Predicting the zoonotic capacity of mammals to transmit SARS-CoV-2

Ilya R. Fischhoff, Adrian A. Castellanos, João P.G.L.M. Rodrigues, Arvind Varsani, Barbara A. Han Proceedings of the Royal Society B DOI: 10.1098/rspb.2021.1651

METHODS

Protein sequence and alignment:

We assembled a dataset of ACE2 NCBI GenBank accessions that are known human ACE2 orthologs or have high similarity to known orthologs as determined using BLASTx (Altschul et al. 1990). Using the R package rentrez and the accession numbers, we downloaded ACE2 protein sequences (Winter 2017). We supplemented these sequences by manually downloading four additional sequences from the MEROPS database (Rawlings et al. 2018).

Structural Modeling of ACE2 orthologs bound to SARS-CoV-2 spike:

Sequences of ACE2 orthologs were aligned using MAFFT (Katoh et al. 2002) and trimmed to the region resolved in the template crystal structure of hACE2 bound to the SARS-CoV-2 spike (PDB ID: 6m0j, Lan et al. 2020). Ambiguous positions in each sequence, artifacts of the sequencing method, were replaced by Glycine to minimize assumptions about the nature of the amino acid side-chain but still allow for modeling. For each ortholog, we generated 10 homology models using MODELLER 9.24 (Sali and Blundell 1993, Webb and Sali 2016), with restricted optimization (fastest schedule) and refinement (very_fast schedule) settings, and selected a representative model based on the normalized DOPE score. These representative models were then manually inspected and 27 were removed from further analysis due to large insertions/deletions or to the presence of too many ambiguous amino acids at the interface with spike. Each validated model was submitted for refinement to the HADDOCK web server (van Zundert et al. 2016), which ran 50 independent short molecular dynamics simulations in explicit solvent to optimize the interface between the two proteins. For each one of the animal species in our study, we assigned an average and standard deviation of the scores of the 10 best refined models, ranked by their HADDOCK score -- a combination of van der Waals, electrostatics, and desolvation energies. Code used for these analyses is freely available (https://zenodo.org/record/4517509).

Trait data collection, processing, and transformations:

We initially created a dataset of vertebrate traits to take advantage of the broad taxonomic breadth of ACE2 sequences available. However, since all known species susceptible to SARS-CoV-2 are mammals, the primary focus of this study is on the modeling and prediction of susceptibility to SARS-CoV-2 in mammals, with vertebrate models included to show our

progression through these data. Vertebrate analyses are not discussed in the main manuscript. For the mammal-specific analyses, we added additional trait variables compiled for mammals (Supplementary Table 2). While we did not pursue the vertebrate modeling process further, information about how the vertebrate dataset was created, and how initial models were evaluated is reported here. The descriptions below describe the overall process of trait data collection and processing.

As many trait databases use older taxonomic references (e.g., the third edition of Mammal Species of the World [Wilson and Reeder 2005], Birdlife V3 world list as modified by Jetz et al. 2012), species names were reverted to older synonyms associated with trait data when appropriate. We primarily sourced trait data from AnAge (de Magalhães and Costa 2009), Amniote Life History Database (Myhrvold et al., 2015), and EltonTraits (Wilman et al., 2014) (Supplementary Table 2), and augmented these data with values from AmphiBIO (Oliveira et al., 2017), FishBase (Froese and Pauly, 2019), (Meiri, 2018), (Meiri, 2019), and a database of CITES listed turtles and tortoises ("Open Data," n.d.) to increase variable coverage across our species.

We also engineered additional traits that have shown importance in predicting host pathogen associations in other contexts (e.g., habitat breadth using the percentage of ecoregions a species occupies; Dallas et al. 2017). We used the sf R package (Pebesma, 2018) to calculate range size from IUCN RedList (IUCN, 2020) and Birdlife International species range polygons ("BirdLife Data Zone," 2021) and Food and Agriculture Organization of the United Nations (FAO) polygons of Major Fishing Areas ("FAO Fisheries & Aquaculture," 2020). To determine overlap with ecoregions, we used polygons of terrestrial ecoregions from The Nature Conservancy (Olson and Dinerstein, 2002) and marine ecoregions from the Marine Ecoregions of the World by the World Wildlife Fund (Spalding et al., 2007). These polygons were rasterized with a resolution of ~100 km² using the R package fasterize (Ross, 2020). We counted the number of unique ecoregions each species overlapped. To enable comparison across species, we expressed ecoregion breadth as the percentage of all terrestrial ecoregions (for terrestrial and freshwater species), marine ecoregions (for marine species according to FAO code), or both (for species found in both marine and freshwater environments according to their FAO Major Fishing Area). Given the coarseness of fish polygon data, we instead gathered georeferenced point data for all fishes in our dataset from the Global Biodiversity Information Facility (DOI: 10.15468/DL.X55PZR) and checked for ecoregion overlap with these points. Mass specific production as described by (Hamilton et al., 2011) was calculated using adult body mass, birth/hatching weight, litter/clutch size, and number of litters/clutches per year.

To account for potential effects on our models related to study effort, we used the wosr package (Baker, 2018) to determine the number of publications with each species name in their title or abstract in Web of Science. As this citation count could itself be biased towards species with recent name changes or a high degree of taxonomic uncertainty, we determined synonyms for each of our species using the backbone taxonomy from GBIF (GBIF Secretariat, 2019) and used a species' name and all its synonyms to craft our queries.

Following the results of initial structural modeling (described above), we observed that perresidue energy decomposition analysis of HADDOCK scores for 29 species indicated that all species with strong predicted binding had in common a salt bridge between SARS-CoV-2 K417 and a negatively charged amino acid at position 30 in the ACE2 sequence (Rodrigues et al. 2020). Given the apparent effect of amino acid 30 on overall binding strength, we constructed an additional feature to denote whether amino acid 30 is negatively charged (and therefore more likely to support strong binding) and included this feature as an additional trait in our models. We used a PROMALS3D (Pei et al., 2008) alignment of our ACE2 sequences to extract residue identity at position 30 and encode whether the residue at this position was negatively charged.

To compare foraging strategy variables across birds and mammals and make these variables consistent across classes, we reclassified them from a continuous percentage to a binary variable. The bird specific EltonTraits variables of ForStrat-watbelowsurf and ForStrat-wataroundsurf were reclassified as ForStrat-ground to match with the definition given for this variable in the mammal EltonTraits data. Additionally, ForStrat-midhigh and ForStrat-canopy were merged with ForStrat-arboreal and ForStrat-pelagic with ForStrat-marine. For fishes, FAO Major Fishing Areas referencing inland and oceanic waters were used to determine values in ForStrat-ground (non-marine aquatic) and ForStrat-marine. As diet categories for fishes and terrestrial vertebrates did not match, the number of nonzero values in diet categories (e.g., fruit, seeds, invertebrates) for a species were counted and then divided by the total number of terrestrial or marine diet categories to enable comparison across classes. For any species without mean longevity information from the Amniote Life History Database, we added values for maximum longevity if they existed. Incubation and gestation time showed a bias in coverage related to class, so we merged the two variables, taking the mean in the event that values for both variables existed. We log transformed values for any variable with a heavily skewed distribution as shown by a density curve. For trait sources, units, coverage, and transformation information, see Supplementary Table 2.

For prediction and imputation of missing values, the same trait variables mentioned in the main text for our mammal-only dataset (PanTHERIA traits, those from Amniote Life History Database, AnAge, EltonTraits, and those calculated for this study; traits with values reported in the "Coverage mammal (%)" column of Supplementary Table 2) were compiled for the 5,400 mammal species in the EltonTraits dataset. We chose this set of species because they followed the taxonomy of the 2005 edition of Mammal Species of the World (Wilson and Reeder, 2005) and showed high coverage across their variables, making this dataset an ideal starting point to add variables from other trait databases.

Given that cetaceans are the main taxonomic group in mammals with marine foraging strategies, we adjusted the ForStrat variables used with our mammal-only dataset to reduce possible bias. We condensed our foraging strategy variables into terrestrial and aquatic versions, with aquatic foraging strategies including those freshwater foraging species previously classified as ForStrat.ground (see Supplementary Table 2).

Imputation of missing mammal trait values

To boost the predictive power of our mammal dataset, we imputed missing values of variables. For variables with more than 1% and less than 95% coverage in the dataset of 5,400 mammals, we used boosted regression tree modeling to model each variable as a Gaussian outcome. As a relatively large sample size was typically present for training these models, parameterization explored learning rates of 0.001 and 0.01, maximum interaction depth of 1, 2, 3, and 4, and a

minimum number of observations in the terminal nodes of 2, 5, and 10. Combinations of parameters were compared using pseudo- R^2 . Those parameter combinations with the best performance were then evaluated through 10 rounds of bootstrapping. For this model evaluation, bootstrapping used different seeds to partition training and test datasets for each iteration (for more detailed information on the bootstrapping process we used, see the "Bootstrapping to describe precision and accuracy of models and predictions" subsection below). After model evaluation with all predictors and those with importance over one percent, we applied two filters to determine which variables to impute missing values for: 1) corrected mean test pseudo- R^2 over 0.75 for either set of bootstrap runs, and 2) correlation over 0.8 between original values and mean values predicted by these same bootstrap iterations. We built models of the variables that passed these filters using the full dataset and used them to predict missing values in our mammal specific dataset.

A small number of imputed data values were biologically impossible: 107 data points (0.6% of all data), across 3 variables (female_maturity_d, mass_specific_production, and longevity_y) had negative values. We retained these values to avoid introducing bias associated with constraining values (Rodwell et al., 2014).

Domesticated mammals

Many of the mammals for which we found the strongest evidence of zoonotic capacity are domesticated to some degree (pets, farmed or traded animals, lab models; Oude Munnink et al. 2020, Schlottau et al. 2020, Shi et al. 2020). Relative to their ancestors or wild conspecifics, domesticated animals often have distinctive traits (Wilkins et al. 2014) that are likely to influence the number of zoonoses found in these species (Cleaveland et al. 2001). To account for trait variation due to domestication in certain species, we modeled mammals in two ways. First, we incorporated a variable indicating whether the source populations from which trait data were collected are wild or non-wild (e.g., farmed, pets, laboratory animals; non-wild status confirmed by the Mammal Diversity Database [Mammal Diversity Database 2020]). Trait data collected from both wild and non-wild individuals were considered to represent non-wild species for the purposes of this model. In a second approach, we used only the wild species for model training and evaluation. The latter approach resulted in higher model accuracy, thus we used this approach in making predictions for mammals. For both approaches, pre-imputation trait values were used for all non-wild mammals during model training, evaluation, and prediction.

Modeling:

We used generalized boosted regression to generate predictions using species' trait data as predictors of the following labels: 1) HADDOCK score (with a continuous distribution), 2) zoonotic capacity (a binary label based on a HADDOCK score threshold, described in detail below under Modeling), and 3) the charge at amino acid position 30 (a binary label, also described below under Modeling). Generalized boosted regression is an ensemble machine learning approach that accommodates non-random patterns of missing data, nonlinear relationships and interacting effects among predictors. In a boosted regression model, a sequence of regression models are fit by recursive binary splits, with each additional regression modelling

those instances that were poorly accounted for by the previous regression iterations in the tree (Elith et al. 2008). Although we created models using both vertebrate and mammal-only datasets, this study focuses on the results and predictions made from mammal-only models (Supplementary Table 3). Variables used for each model can be seen in Supplementary Table 2. Before modeling, we removed variables with near zero variation using the nearZeroVar function in the caret R package (Kuhn 2020). All variables with near zero variation were ordinal level taxonomic variables (specifically, Afrosoricida, Artiodactyla, Cingulata, Dasyuromorphia, Didelphimorphia, Diprotodontia, Erinaceomorpha, Lagomorpha, Macroscelidea, Monotremata, Perissodactyla, Pholidota, Proboscidea, Sirenia, Soricomorpha, and Tubulidentata).

Grid search for selecting hyperparameters

For all models, we applied a grid search to select optimal hyperparameters. This parameter search included fitting models with all possible combinations of learning rates of 0.0001, 0.001, 0.001, and 0.1, maximum interaction depths among variables of 2, 3, and 4, and minimum number of observations in a terminal node in the decision tree (n.minobsinnode) of 2, 3, 4, or 5. We chose values for the minimum number of observations in terminal nodes smaller than the default of 10 due to the small size of the dataset. Each model included a maximum of 100,000 trees. For each set of hyperparameters, we graphed the training and test deviance by number of trees. For some hyperparameters, the optimal number of trees, with the lowest test deviance, was reached rapidly, followed by rapid increase in deviance with increasing number of trees. Based on visual inspection, we selected the set of hyperparameters that had the highest test accuracy and stably decreasing deviance curves.

We produced three generalized boosted regression models. For the zoonotic capacity model (the main model described in this paper) and the amino acid 30 charge model, a Bernoulli error distribution was used for binary classification. We measured performance by the area under the receiver operating characteristic curve (AUC) on a hold-out test dataset (20% of the labelled data). For the continuous binding strength model, which used a Gaussian error distribution, performance was measured using pseudo- R^2 .

Modeling

We began by modeling HADDOCK score for all vertebrates using boosted regression. When we found a relatively low pseudo-R² for this model (Supplementary Table 3), we designed two simpler models. One simpler model classified whether a species' HADDOCK score is at or below -129. This value is between two HADDOCK scores: the domestic cat (*Felis catus*), which is currently the species with weakest predicted binding among animals with confirmed conspecific transmission (Bosco-Lauth et al., 2020), and the pig / wild boar (*Sus scrofa*), which shows the strongest predicted binding among species for which experimental inoculation failed to cause detectable infection (Shi et al., 2020). A second simpler model classified whether a species has a negative charge at amino acid residue 30 in ACE2, which appeared to be a useful proxy for susceptibility based on previous computational and empirical results suggesting that a negative charge at this residue is an important shared feature of species with strong binding to SARS-CoV-2 (Rodrigues et al., 2020). We fit amino acid 30 charge models including HADDOCK score as a predictor. Our trait-based models predicted nearly all mammal species to have a negative charge at amino acid 30, thus the results of this model were relatively

uninformative except to identify a few species with particularly low risk of binding SARS-CoV-2. For full model results, see Supplementary Table 3.

Bootstrapping to describe precision and accuracy of models and predictions

Applying the set of hyperparameters selected by grid search, we performed model fitting 50 times for the zoonotic capacity model and 10 times for all other models, using different seeds each time for partitioning training and test datasets. Across these bootstrapping runs, we computed the mean training and test accuracy measure, AUC for binary classification and pseudo- R^2 for models of the continuous HADDOCK score.

To gain a more realistic measure of accuracy, we completed a second set of bootstrapping runs to develop a null distribution of accuracy scores. Patterns in the structure of the predictor data, independent of their relationship to the label, have the potential to influence apparent model accuracy. To account for this possibility in a null model, we randomly permuted labels prior to modeling. We repeated this procedure 50 times (for the zoonotic capacity model) or 10 times (for the other models) to produce a null distribution of accuracy statistics. For binary classification models, we corrected the test AUC by the difference between the null model test AUC mean (across bootstrap runs) and 0.5, the latter being the performance expected by a model that performed no better than chance. For models with Gaussian distribution (for HADDOCK scores), we reached a corrected pseudo- R^2 by subtracting the absolute value of the null model pseudo- R^2 mean.

We used all available data to train one model of zoonotic capacity that we used to make predictions for each of 5,400 mammal species. To develop confidence estimates around these zoonotic capacity predictions, each of the 50 bootstrap iterations were also used to make predictions on the full mammal dataset. Zoonotic capacity results for the full data predictions and 50 bootstrap predictions can be found in Supplementary File 1.

Standardizing species susceptibility predictions across studies:

To place our methods and predictions in context with the variety of existing approaches used to predict species susceptibility to SARS-CoV-2, we collated the results of previous studies that made predictions on multiple animal species. These studies made predictions through sequence-based methods (e.g., comparing species' ACE2 amino acid sequence similarity with human ACE2), structure-based methods (e.g., creating three-dimensional representations of ACE2 orthologs), and laboratory experiments (e.g., characterizing cell entry using pseudotyped viruses). We collated the predictions made by the following studies: Ahmed et al., 2021, Damas et al., 2020, Huang et al., 2020, Kumar et al., 2020, Lam et al., 2020, Li et al., 2020, Y. Liu et al., 2020, Z. Liu et al., 2020, Luan et al., 2020, Mathavarajah et al., 2021, Melin et al., 2020, Rodrigues et al., 2020, Yan et al., 2020, and Zhao et al., 2020.

Some of these studies categorized species susceptibility as a gradient. For instance, Damas et al., 2020 ranked species as a series of five categories from very high to very low, and Mathavarajah et al., 2020 ranked species as high, medium and low. Other studies categorized species

susceptibility as a binary measure. For example, Huang et al., 2020 report whether each species is predicted to be susceptible or not. For those studies that did not assign species into explicit categories, we assigned categories based on the context of conclusions discussed by the authors of each study. For example, Rodrigues et al. 2020 discuss their species predictions by comparing each species' HADDOCK score relative to empirical studies that confirm species as either positive or negative for SARS-CoV-2 infection. Based on the context of this discussion, we binned all of the species predictions from Rodrigues et al. 2020 into high (e.g., orangutan) and low (e.g., hedgehog) categories. For ease of comparison (Figure 1), we used three categories (low, medium, high) to organize predictions of species susceptibility across studies. Damas et al. 2020 was the only study that used categories outside of these. Very low and very high predictions from Damas et al. 2020 were recategorized as low and high, respectively. All other categorizations made by study authors were retained.

For the alluvial plot comparing species predictions by study method (Supplementary Figure 1), all species predictions were transformed from a character value to a numeric value (e.g., low susceptibility translates to a 1 and high susceptibility translates to a 3). For species with multiple predictions generated for a single method, predictions were averaged together and then rounded to the nearest whole number to create a single value for each species-method combination (e.g., predictions of high [3] and low [1] for *Sciurus niger* from sequence-based prediction would be translated to an overall prediction of medium [2] for this species-method pairing).

Mapping:

To identify areas for future surveillance work, we mapped the top 10% of predictions from the zoonotic capacity threshold model (min probability = 0.748, mean = 0.889, max = 0.988). For representations of each species' distribution, we used species range polygons from the IUCN Red List (IUCN 2020). Additionally, we focused on species found in human occupied or human modified landscapes because co-occurrence with humans may increase risk of spillback or secondary spillover. We used the IUCN API and the rredlist package in R version 4.0.0 to gather habitat associations for our subset of species predictions (Chamberlain, 2020; IUCN, 2020; R Core Team, 2020), and subset to species listed by IUCN as having artificial terrestrial habitat associations (e.g., urban areas, crops or pastures, heavily degraded landscapes, etc.). We filtered these species distributions to only those areas overlapping with areas classified as 'artificial terrestrial' (according to classifications created by Jung et al., 2020).

Due to known issues matching high resolution environmental data with the much coarser resolution of range polygons (see Jetz et al., 2012 for a good explanation of the issue), we resampled the artificial terrestrial habitat raster from ~1 km² to ~100 km² in resolution using the resample function in the raster package in R (Hijmans, 2020). Polygons of species distribution data were rasterized at a resolution of ~100 km² to match these data using the fasterize package in R (Ross, 2020). Small polygons may not be rasterized correctly using this method, so any polygon that produced an empty raster was converted to point data and rasterized using the rasterize function in the raster package (Hijmans, 2020). Any cells for rasterized species distributions that did not overlap a cell classified as artificial terrestrial were changed to NA.

Regions with higher numbers of human COVID-19 cases may pose a higher risk of spillback transmission to co-occurring animals. We gathered cumulative case count data from the COVID-

19 Data Repository by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University (Dong et al., 2020). These data are mostly at the country level with the exception of state or province information for Australia, Brazil, Canada, Chile, Colombia, France, Germany, India, Italy, Japan, Mexico, the Netherlands, Pakistan, Peru, Russia, Spain, Sweden, Ukraine, the United Kingdom, and the United States of America. We considered hotspots as those administrative units (i.e., countries or states/provinces) with 100,000 or more cumulative COVID-19 cases. Species distributions were further restricted to areas within regions designated as hotspots to determine a final map of highest spillback risk (panel C of Figure 4 in the main text).

References:

- Ahmed R, Hasan R, Siddiki AMAMZ, Islam MS. 2021. Host range projection of SARS-CoV-2: South Asia perspective. *Infect Genet Evol* **87**:104670.
- Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990 Basic local alignment search tool. J. Mol. Biol. 215, 403–410. (doi:10.1016/S0022-2836(05)80360-2)
- Baker C. 2018. wosr: Clients to the "Web of Science" and "InCites" APIs.

BirdLife Data Zone. 2021.

- Bosco-Lauth AM, Hartwig AE, Porter SM, Gordy PW, Nehring M, Byas AD, VandeWoude S, Ragan IK, Maison RM, Bowen RA. 2020. Experimental infection of domestic dogs and cats with SARS-CoV-2: Pathogenesis, transmission, and response to reexposure in cats. *Proc Natl Acad Sci U S A* 117:26382–26388.
- Chamberlain S. 2020. rredlist: IUCN Red List Client. https://CRAN.Rproject.org/package=rredlist
- Cleaveland S, Laurenson MK, Taylor LH. 2001 Diseases of humans and their domestic mammals: pathogen characteristics, host range and the risk of emergence. Philos. Trans. R. Soc. Lond. B Biol. Sci. 356, 991–999. (doi:10.1098/rstb.2001.0889)
- Dallas T, Park AW, Drake JM. 2017 Predicting cryptic links in host-parasite networks. PLoS Comput. Biol. 13, e1005557. (doi:10.1371/journal.pcbi.1005557)
- Damas J, Hughes GM, Keough KC, Painter CA, Persky NS, Corbo M, Hiller M, Koepfli K-P, Pfenning AR, Zhao H, Genereux DP, Swofford R, Pollard KS, Ryder OA, Nweeia MT, Lindblad-Toh K, Teeling EC, Karlsson EK, Lewin HA. 2020. Broad host range of SARS-CoV-2 predicted by comparative and structural analysis of ACE2 in vertebrates. *Proc Natl Acad Sci U S A* 117:22311–22322.
- de Magalhães JP, Costa J. 2009. A database of vertebrate longevity records and their relation to other life-history traits. *J Evol Biol* **22**:1770–1774.
- Dong E, Du H, Gardner L. 2020. An interactive web-based dashboard to track COVID-19 in real time. *Lancet Infect Dis* **20**:533–534.

- Elith, J, Leathwick JR, Hastie T. 2008. A working guide to boosted regression trees. *J Anim Ecol* **77**, 802-813.
- FAO Fisheries & Aquaculture. 2020. http://www.fao.org/fishery/area/search/en
- Froese R, Pauly D. 2019. FishBase. Fishbase. https://www.fishbase.de/
- GBIF Secretariat. 2019. GBIF Backbone Taxonomy. doi:10.15468/39omei
- Hamilton MJ, Davidson AD, Sibly RM, Brown JH. 2011. Universal scaling of production rates across mammalian lineages. *Proceedings of the Royal Society B: Biological Sciences* 278:560–566.
- Hijmans RJ. 2020. Geographic Data Analysis and Modeling [R package raster version 3.4-5].
- Huang X, Zhang C, Pearce R, Omenn GS, Zhang Y. 2020. Identifying the Zoonotic Origin of SARS-CoV-2 by Modeling the Binding Affinity between the Spike Receptor-Binding Domain and Host ACE2. J Proteome Res 19:4844–4856.
- IUCN. 2020. The IUCN Red List of Threatened Species.
- Jetz W, McPherson JM, Guralnick RP. 2012. Integrating biodiversity distribution knowledge: toward a global map of life. *Trends Ecol Evol* **27**:151–159.
- Jetz W, Thomas GH, Joy JB, Hartmann K, Mooers AO. 2012. The global diversity of birds in space and time. *Nature* **491**:444–448.
- Jung M, Dahal PR, Butchart SHM, Donald PF, De Lamo X, Lesiv M, Kapos V, Rondinini C, Visconti P. 2020. A global map of terrestrial habitat types. *Sci Data* **7**:256.
- Katoh K, Misawa K, Kuma K-I, Miyata T. 2002 MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res. 30, 3059–3066. (doi:10.1093/nar/gkf436)
- Kuhn M. 2020. caret: Classification and Regression Training. R package version 6.0-86. https://CRAN.R-project.org/package=caret
- Kumar A, Pandey SN, Pareek V, Narayan RK, Faiq MA, Kumari C. 2020. Predicting susceptibility for SARS-CoV-2 infection in domestic and wildlife animals using ACE2 protein sequence homology. *Zoo Biol.* doi:10.1002/zoo.21576
- Lam SD, Bordin N, Waman VP, Scholes HM, Ashford P, Sen N, van Dorp L, Rauer C, Dawson NL, Pang CSM, Abbasian M, Sillitoe I, Edwards SJL, Fraternali F, Lees JG, Santini JM, Orengo CA. 2020. SARS-CoV-2 spike protein predicted to form complexes with host receptor protein orthologues from a broad range of mammals. *Sci Rep* 10:16471.
- Lan J et al. 2020 Structure of the SARS-CoV-2 spike receptor-binding domain bound to the ACE2 receptor. Nature 581, 215–220. (doi:10.1038/s41586-020-2180-5)

- Liu Y, Hu G, Wang Y, Zhao X, Ji F, Ren W, Gong M, Ju X, Li C, Hong J, Zhu Y, Cai X, Wu J, Lan X, Xie Y, Wang X, Yuan Z, Zhang R, Ding Q. 2020. Functional and Genetic Analysis of Viral Receptor ACE2 Orthologs Reveals a Broad Potential Host Range of SARS-CoV-2. *bioRxiv*. doi:10.1101/2020.04.22.046565
- Liu Z, Xiao X, Wei X, Li J, Yang J, Tan H, Zhu J, Zhang Q, Wu J, Liu L. 2020. Composition and divergence of coronavirus spike proteins and host ACE2 receptors predict potential intermediate hosts of SARS-CoV-2. *J Med Virol* **92**:595–601.
- Li Y, Wang H, Tang X, Fang S, Ma D, Du C, Wang Y, Pan H, Yao W, Zhang R, Zou X, Zheng J, Xu L, Farzan M, Zhong G. 2020. SARS-CoV-2 and Three Related Coronaviruses Utilize Multiple ACE2 Orthologs and Are Potently Blocked by an Improved ACE2-Ig. *J Virol* 94. doi:10.1128/JVI.01283-20
- Luan J, Jin X, Lu Y, Zhang L. 2020. SARS-CoV-2 spike protein favors ACE2 from Bovidae and Cricetidae. *J Med Virol* **92**:1649–1656.
- Mammal Diversity Database. 2020 Mammal Diversity Database. (doi:10.5281/zenodo.4139818)
- Mathavarajah S, Stoddart AK, Gagnon GA, Dellaire G. 2021. Pandemic danger to the deep: The risk of marine mammals contracting SARS-CoV-2 from wastewater. *Sci Total Environ* **760**:143346.
- Meiri S. 2019. Endothermy, offspring size and evolution of parental provisioning in vertebrates. *Biol J Linn Soc Lond* **128**:1052–1056.
- Meiri S. 2018. Traits of lizards of the world: Variation around a successful evolutionary design: Meiri. *Glob Ecol Biogeogr* **27**:1168–1172.
- Melin AD, Janiak MC, Marrone F 3rd, Arora PS, Higham JP. 2020. Comparative ACE2 variation and primate COVID-19 risk. *Commun Biol* **3**:641.
- Myhrvold NP, Baldridge E, Chan B, Sivam D, Freeman DL, Ernest SKM. 2015. An amniote life-history database to perform comparative analyses with birds, mammals, and reptiles: Ecological ArchivesE096-269. *Ecology* **96**:3109–3000.
- Oliveira BF, São-Pedro VA, Santos-Barrera G, Penone C, Costa GC. 2017. AmphiBIO, a global database for amphibian ecological traits. *Sci Data* **4**:170123.
- Olson DM, Dinerstein E. 2002. The Global 200: Priority Ecoregions for Global Conservation. *Ann Mo Bot Gard* **89**:199–224.
- Open Data. n.d. https://www.species360.org/serving-conservation/turtles-tortoises-cites/
- Oude Munnink BB et al. 2020 Transmission of SARS-CoV-2 on mink farms between humans and mink and back to humans. Science (doi:10.1126/science.abe5901)

- Pebesma E. 2018. Simple Features for R: Standardized Support for Spatial Vector Data. *R J* **10**:439–446.
- Pei J, Kim B-H, Grishin NV. 2008. PROMALS3D: a tool for multiple protein sequence and structure alignments. *Nucleic Acids Res* **36**:2295–2300.
- R Core Team. 2020. R: A language and environment for statistical computing. Vienna, Austria.
- Rawlings ND, Barrett AJ, Thomas PD, Huang X, Bateman A, Finn RD. 2018 The MEROPS database of proteolytic enzymes, their substrates and inhibitors in 2017 and a comparison with peptidases in the PANTHER database. Nucleic Acids Res. 46, D624–D632. (doi:10.1093/nar/gkx1134)
- Rodrigues JPGLM, Barrera-Vilarmau S, M C Teixeira J, Sorokina M, Seckel E, Kastritis PL, Levitt M. 2020. Insights on cross-species transmission of SARS-CoV-2 from structural modeling. *PLoS Comput Biol* 16:e1008449.
- Rodwell L, Lee KJ, Romaniuk H, Carlin JB. 2014. Comparison of methods for imputing limitedrange variables: a simulation study. *BMC Med Res Methodol* 14:57.
- Ross N. 2020. Fast Polygon to Raster Conversion. Comprehensive R Archive Network (CRAN).
- Sali A, Blundell TL. 1993 Comparative protein modelling by satisfaction of spatial restraints. J. Mol. Biol. 234, 779–815. (doi:10.1006/jmbi.1993.1626)
- Schlottau K et al. 2020 SARS-CoV-2 in fruit bats, ferrets, pigs, and chickens: an experimental transmission study. Lancet Microbe 1, e218–e225. (doi:10.1016/S2666-5247(20)30089-6)
- Shi J, Wen Z, Zhong G, Yang H, Wang C, Huang B, Liu R, He X, Shuai L, Sun Z, Zhao Y, Liu P, Liang L, Cui P, Wang J, Zhang X, Guan Y, Tan W, Wu G, Chen H, Bu Z. 2020. Susceptibility of ferrets, cats, dogs, and other domesticated animals to SARS-coronavirus 2. *Science* 368:1016–1020.
- Spalding MD, Fox HE, Allen GR, Davidson N, Ferdaña ZA, Finlayson M, Halpern BS, Jorge MA, Lombana A, Lourie SA, Martin KD, McManus E, Molnar J, Recchia CA, Robertson J. 2007. Marine Ecoregions of the World: A Bioregionalization of Coastal and Shelf Areas. *Bioscience* 57:573–583.
- van Zundert GCP et al. 2016 The HADDOCK2.2 Web Server: User-Friendly Integrative Modeling of Biomolecular Complexes. J. Mol. Biol. 428, 720–725. (doi:10.1016/j.jmb.2015.09.014)
- Webb B, Sali A. 2016 Comparative Protein Structure Modeling Using MODELLER. Curr. Protoc. Bioinformatics 54, 5.6.1–5.6.37. (doi:10.1002/cpbi.3)
- Wilkins AS, Wrangham RW, Fitch WT. 2014 The 'domestication syndrome' in mammals: a unified explanation based on neural crest cell behavior and genetics. Genetics 197, 795–808. (doi:10.1534/genetics.114.165423)

- Wilman H, Belmaker J, Simpson J, de la Rosa C, Rivadeneira MM, Jetz W. 2014. EltonTraits 1.0: Species-level foraging attributes of the world's birds and mammals: Ecological ArchivesE095-178. *Ecology* 95:2027–2027.
- Wilson DE, Reeder DM. 2005. Mammal Species of the World: A Taxonomic and Geographic Reference. Baltimore, MD: JHU Press.
- Winter D. 2017 rentrez: An R package for the NCBI eUtils API. R J. 9, 520. (doi:10.32614/rj-2017-058)
- Yan H, Jiao H, Liu Q, Zhang Z, Wang X, Guo M, Wang B-J, Lan K, Chen Y, Zhao H. 2020. Many bat species are not potential hosts of SARS-CoV and SARS-CoV-2: Evidence from ACE2 receptor usage. *bioRxiv*. doi:10.1101/2020.09.08.284737
- Zhao X, Chen D, Szabla R, Zheng M, Li G, Du P, Zheng S, Li X, Song C, Li R, Guo J-T, Junop M, Zeng H, Lin H. 2020. Broad and Differential Animal Angiotensin-Converting Enzyme 2 Receptor Usage by SARS-CoV-2. *J Virol* **94**. doi:10.1128/JVI.00940-20