SARS-CoV-2 spike protein predicted to form stable complexes with host receptor protein orthologues from mammals, but not fish, birds or reptiles

Lam SD[1,2,†], Bordin N[2,†], Waman VP[2], Scholes HM[2], Ashford P[2], Sen N[2,3], van Dorp L[4], Rauer C[2], Dawson NL[2], Pang CSM[2], Abbasian M[2], Sillitoe I[2], Edwards SJL[5], Fraternali F[6], Lees JG[7], Santini JM[2], Orengo CA[2,*]

1 Department of Applied Physics, Faculty of Science and Technology, Universiti Kebangsaan Malaysia, Bangi, Selangor, 43600, Malaysia

2 Institute of Structural and Molecular Biology, University College London, London, WC1E 6BT, UK

3 Indian Institute of Science Education and Research, Pune, 411008, India

4 UCL Genetics Institute, University College London, London, WC1E 6BT, UK

5 Department of Science and Technology Studies, University College London, London, WC1E 6BT, UK

6 Randall Division of Cell and Molecular Biophysics, Guy's Campus, New Hunt's House, King's College London, London, SE1 1UL, UK

7 Department of Biological and Medical Sciences, Faculty of Health and Life Sciences, Oxford Brookes University, Oxford, OX3 OBP, UK

† These authors contributed equally

* To whom correspondence should be addressed: c.orengo@ucl.ac.uk

Abstract

The coronavirus disease 2019 (COVID-19) global pandemic is caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). SARS-CoV-2 has a zoonotic origin and was transmitted to humans via an undetermined intermediate host, leading to widespread infections in humans and reports of infections in other mammals. To enter host cells, the viral spike protein binds to its receptor, angiotensin-converting enzyme 2 (ACE2), and is processed by a protease, transmembrane protease serine 2 (TMPRSS2). Whilst receptor binding contributes to the viral host range, changes in energy of the spike protein:ACE2 complexes in other animals have not been widely explored. Here, we analyse interactions between the spike protein and orthologues of ACE2 and TMPRSS2 from a broad range of 215 vertebrate species. Structures of these orthologues have not been determined, so we predicted them. Using models of the spike protein:ACE2 orthologue complexes, we calculated their changes in energy, and correlated these to COVID-19 severities in mammals. Across vertebrate orthologues, mutations are predicted to be more disruptive to the structure of ACE2 than TMPRSS2. Finally, we provide phylogenetic evidence that SARS-CoV-2 has recently transmitted from humans to animals. Our results suggest that SARS-CoV-2 can infect a broad range of mammals—but not fish, birds or reptiles—which could serve as reservoirs of the virus, necessitating careful ongoing animal management and surveillance.

Keywords: SARS-CoV-2, COVID-19, spike protein, ACE2, TMPRSS2, structural bioinformatics

Introduction

Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) is a novel coronavirus that emerged towards the end of 2019 and is responsible for the coronavirus disease 2019 (COVID-19) global pandemic. Available data suggests that SARS-CoV-2 has a zoonotic source[1], with the closest sequence currently available deriving from the horseshoe bat[2]. As yet the transmission route to humans, including the intermediate host, is unknown. So far, little work has been done to assess the animal reservoirs of SARS-CoV-2, or the potential for the virus to spread to other species living with, or in close proximity to, humans in domestic, rural, agricultural or zoological settings.

Coronaviruses, including SARS-CoV-2, are major multi-host pathogens and can infect a wide range of non-human animals[3–5]. SARS-CoV-2 falls within the betacoronavirus genus, which includes species which infect economically important livestock including cows[6] and pigs[7], mice, rats[8] and rabbits[9], and wildlife such as antelope and giraffe[10]. Severe acute respiratory syndrome coronavirus (SARS-CoV), the betacoronavirus that caused the 2002-2004 SARS outbreak[11], likely jumped to humans from its original bat host via civets. Viruses genetically similar to human SARS-CoV have been isolated from animals as diverse as civets, racoon dogs, ferret-badgers[3] and pigs[12], suggesting the existence of a large host reservoir. It is therefore probable that SARS-CoV-2 can also infect a wide range of species.

Real-world SARS-CoV-2 infections have been reported in cats[13], tigers[14], dogs[15,16] and mink[15,16]. Animal infection studies have also identified cats[17] and dogs[17] as hosts, as well as ferrets[17], macaques[18] and marmosets[18]. Recent *in vitro* studies have also suggested an even broader set of animals may be infected[19–21]. As with humans, age is also a predictive risk factor for rhesus macaques, with SARS-CoV-2 causing worse pneumonia in older animals[22]. To understand the potential host range of SARS-CoV-2, the plausible extent of zoonotic and anthroponotic transmission, and to guide surveillance efforts, it is vital to know which species are susceptible to SARS-CoV-2 infection.

The receptor binding domain (RBD) of the SARS-CoV-2 spike protein (S-protein) binds to the extracellular peptidase domain of angiotensin I converting enzyme 2 (ACE2) mediating cell entry[23]. ACE2 is highly expressed in the endothelium, kidney, lungs and heart[24,25] and plays a role in regulating blood pressure[26], with loss of function increasing the risk of heart failure[27]. The sequence of ACE2 is highly conserved across vertebrates, suggesting that SARS-CoV-2 could use orthologues of ACE2 for cell entry. The structure of the SARS-CoV-2 S-protein RBD has been solved in complex with human ACE2[28] (PDB ID 6M0J) and identification of critical binding residues in this structure have provided valuable insights into viral recognition of the host receptor[28–32]. Compared with SARS-CoV, the SARS-CoV-2 S-protein has a 10-22-fold higher affinity for human ACE2[28,29,33] thought to be mediated by three classes of mutations in the SARS-CoV-2 S-protein[34]. Similarly, variations in human ACE2 have also been found to increase affinity for S-protein[35] receptor binding. These factors may contribute to the host range and infectivity of SARS-CoV-2.

Both SARS-CoV-2 and SARS-CoV additionally require the transmembrane protease, serine 2 (TMPRSS2) to mediate cell entry. Together, ACE2 and TMPRSS2 confer specificity of host cell types that the virus can enter, with the S-protein preferentially binding to cells that co-express both ACE2 and TMPRSS2[25,36]. Upon binding to ACE2, the S-protein is cleaved by TMPRSS2 at two cleavage sites on separate loops, which primes the spike for cell entry[37]. TMPRSS2 has been docked against the SARS-CoV-2 S-protein, which revealed its binding site to be adjacent to these two cleavage sites[36]. An approved TMPRSS2 protease inhibitor drug is able to block SARS-CoV-2 cell entry[38], which

demonstrates the key role of TMPRSS2 alongside ACE2. Furthermore, variations in TMPRSS2, as opposed to ACE2, have been suggested as associated with disease severity in COVID-19 patients[39]. TMPRSS2 is frequently mutated in prostate cancer, which may contribute to the higher rate of COVID-19 mortality observed in men compared to women of the same age[40–42]. As such, both ACE2 and TMPRSS2 represent attractive therapeutic targets against SARS-CoV-2.

The structural interplay between ACE2 and the SARS-CoV-2 spike protein has been used to predict possible hosts for SARS-CoV-2. A broad range of hosts have already been proposed, covering hundreds of mammalian species, including tens of bat species[31]; and all classes of vertebrates[43,44], including agricultural species of cow, sheep, goat, bison and water buffalo. In addition, residues in ACE2 have been identified as under positive selection in bats, particularly in regions involved in binding the S-protein[31]. The impacts of mutations in ACE2 orthologues have also been tested, for example structural modelling of ACE2 from 27 primate species[44], demonstrated that apes and African and Asian monkeys may also be susceptible to SARS-CoV-2. However, whilst cell entry is necessary for viral infection, it may not be sufficient alone to cause disease. For example, variations in other proteins may prevent downstream events that are required for viral replication in a new host. Hence, examples of real-world infections[13–16] and experimental data from animal infection studies[17–21] are required to validate hosts that are predicted to be susceptible.

Here, we analysed the effect of known mutations in orthologues of ACE2 and TMPRSS2 from a broad range of 215 vertebrate species, including primates, rodents and other placental mammals; birds; reptiles; and fish. For each species, we generated a 3D model of the ACE2 protein structure from its protein sequence and calculated the impacts of known mutations in ACE2 on the stability of the S-protein:ACE2 complex. We correlated changes in the energy of the complex with changes in the structure of ACE2, chemical properties of the binding interface, and experimental infection phenotypes from *in vivo* and *in vitro* animal studies. To further test our predictions, we performed detailed manual structural analyses, presented as a variety of case studies for different species. Unlike other studies that analyse interactions that the S-protein makes with the host, we also analyse the impact of mutations in vertebrate orthologues of TMPRSS2. Our results suggest that SARS-CoV-2 can infect a broad range of vertebrates, which could serve as reservoirs of the virus, supporting future anthroponotic and zoonotic transmission.

Data

Sequence Data

ACE2 protein sequences for 239 vertebrates, including humans, were obtained from ENSEMBL version 99 and eight sequences from UniProt release 2020_1 (Supplementary Table S1).

TMPRSS2 protein sequences for 278 vertebrate sequences, including the human sequence, were obtained from ENSEMBL (Supplementary Table S2).

A phylogenetic tree of species, to indicate the evolutionary relationships between animals, was downloaded from ENSEMBL

Structural Data

The structure[28] of the SARS-CoV-2 S-protein bound to human ACE2 at 2.45Å was used throughout (PDB ID 6M0J).

Methods

Analyses of sequence similarity between human ACE2 and other vertebrate species

Vertebrate protein sequences were aligned to each other pairwise, and to the human ACE2 sequence, using BLASTp[45].

Domain family detection, alignment construction and conservation of residues in ACE2 and TMPRSS2

In order to detect highly conserved residues likely to be implicated in the domain function (i.e. ACE2 binding residues, TMPRSS2 binding residues in the active site) we mapped the domain sequence to the appropriate functional family (FunFam) in the CATH classification[46]. FunFams are clusters of evolutionary related domain sequences, predicted to have highly similar structures and functions. They have been previously used to analyse the impact of genetic variation on protein function in the context of antibiotic resistance[47] phenotypes and cancer[48] progression.

The ACE2 (or TMPRSS2) sequences were scanned against the seed FunFam hidden Markov model library for the next release of CATH-Gene3D[49] v4.3 (unpublished data) with the best matches resolved by cath-resolve-hits[50] using a bit score cut-off of 25 and a minimum query coverage of 80%. Each hit was subsequently re-aligned to the matching FunFam using Clustal[51] in Jalview[52].

The conservation scores for residues in the ACE2 and TMPRSS2 domains were obtained using ScoreCons[53]. Information contents of FunFam multiple sequence alignments were calculated as the diversity of positions (DOPs) score. Residues are considered highly conserved for DOPs ≥ 70 and ScoreCons ≥ 0.7.

Structure analyses

Identifying residues in ACE2

In addition to residues in ACE2 that contact the S-protein directly, various other studies have considered other residues that are in the second shell, or are buried, and could influence binding[54]. Therefore, in our analyses we built on these approaches and extended them and we have used the following sets:

1. *Direct contact (DC)* residues. This includes a total of 20 residues that are involved in directly contacting interfaces[28] identified by PDBe[57] and PDBSum[55].
2. *Direct Contact Extended (DCEX) residues.* This dataset includes residues within 5Å of binding (DC) residues and other more distant residues within 8Å of DC residues that are likely to be important for binding. These were selected by detailed manual inspection of the complex, and also considering the following criteria: (i) reported evidence from deep mutagenesis[54], (ii) *in silico* alanine scanning (using mCSM-PPI[56]), (iii) residues with high evolutionary conservation patterns identified by the FunFams-based protocol described above i.e. residues identified

with DOPS ≥ 70 and ScoreCons score ≥ 0.7, (iv) allosteric site prediction and (v) sites under positive selection.

To identify residues involved in the S-protein:ACE2 interface, we extracted information from a range of sources including PDBe[57], PDBsum[55], reported structural evidence (such as from crystallography, cryo-EM and homology modelling) in the literature[20,28,36,44,54,57], and manual inspection of the human complex (PDB ID 6M0J) using Chimera[58] (Supplementary Figure S1). Other studies identifying residues included alanine scanning mutagenesis and deep mutagenesis experiments[28,44,57,59,60] .

In addition, in order to identify potential allosteric residues of DCEX residues in ACE2, three different allosteric prediction methods were used namely AlloSitePro[61], ENM[62] and PARS[63]. The protein structure PDB ID 6M0J[28] was used as input, which contains information on ligand binding sites for zinc, chlorine, and NAG. PARS is a normal mode analysis method which calculates normal modes in the presence and absence of a simulated allosteric modulator. A site is predicted as allosteric, if the motions are significantly different. ENM identifies hinge residues as potential functional residues, and relies exclusively on inter-residue contact topology. It uses features of spatial clustering properties and relative solvent accessibility to identify functionally important hinge residues. AlloSitePro, uses both pocket-based structure features with normal mode analyses. The performance of AllositePro has been endorsed by recent studies[64].

We also identified DC and DCEX sites under positive selection using codon-based methods, including mixed effect model of evolution[65], available at the Datamonkey Adaptive Evolution web-server[66]. These methods estimate dN/dS ratio for every codon in an alignment.

Generating 3D-structure models

ACE2

Using the ACE2 protein sequence from each species, structural models were generated for the S-protein:ACE2 complex for 247 animals using the FunMod modelling pipeline[52,53], based on MODELLER[54]. Given that the S-protein has the same sequence in PDB ID 6MOJ, and ACE2 is conserved in vertebrates, models are expected to be high-quality. Models were refined by MODELLER to optimise the geometry of the complex and the interface. Only high-quality models were used in this analysis, with nDOPE[55] score < -1 and with < 10 DCEX residues missing. This gave a final dataset of 215 animals for further analysis.

The modelled structures of ACE2 were compared against the human structure (PDB ID 6M0J) and pairwise, against each other, using SSAP[57]. SSAP returns a score in the range 0-100, with identical structures scoring 100.

TMPRSS2

There is currently no solved structure for TMPRSS2. To build a structural model of TMPRSS2, we selected the best TMPRSS2 model template by performing a HMMER[56] search of the human TMPRSS2 sequence (UniProt accession O15393) against the FunFams, identifying four potential structural templates within FunFam 2.40.10.10-FF-2 (best E-value = 3e-96). After modelling with FunMod, the structure with the best score (nDOPE = -0.773) was selected, based on PDB ID 5I25.

We identified the relevant residues in TMPRSS2 by extracting the active site and cleavage site residue information available on the human TMPRSS sequence (UniProt accession O15393). After we modelled the human TMPRSS2, two residues sets were extracted.
Set 1: Direct Contact (ASCS) residues comprising the active site (three residues) and the cleavage site (2 residues)

Set 2: Direct Contact Extended (ASCSEX) residues: This dataset includes the residues from Set 1 and additional residues from the active site 5Å shell (17 residues), and more distant 8Å shell (23 residues) and the cleavage site 5Å shell (10 residues) and more distant 8Å shell (nine residues). Set 2 contains therefore 64 residues, out of which 21 are residues with high (≥ 70).

Measuring changes in the energy of the S-protein:ACE2 complex in SARS-CoV-2 and SARS-CoV

Since we only have a protein structure for human S-protein:ACE2 we modelled the structure of all the other animal species using established protocols[67,68]. We observed very high conservation of ACE2 amongst animals, with most species sharing 60% or more identical residues with human ACE2. We were therefore confident of obtaining high quality models in which the binding interface would be well constructed. This was assessed using standard measures and we removed any animal species with poor models from the analysis.

We calculated the changes in binding energy of the SARS-CoV-2 S-protein:ACE2 complex and the SARS-CoV S-protein:ACE2 complex of different species, compared to human, following three different protocols:

1. *Protocol 1:* Using the human complex and mutating the residues for the ACE2 interface to those found in the given animal sequence and then calculating the ΔΔG of the complex using both mCSM-PPI1[56] and mCSM-PPI2[69]. This gave a measure of the destabilisation of the complex in the given animal relative to the human complex. Values < 0 are associated with destabilising mutations, whilst values ≥ 0 are associated with stabilising mutations.
2. *Protocol 2:* We repeated the analysis with both mCSM-PPI1 and mCSM-PPI2 as in protocol 1, but using the animal 3D-models instead of ACE2 and calculating the ΔΔG of the complex by mutating the animal ACE2 interface residue to the appropriate human residue. Values > 0 are associated with destabilising mutations, whilst values ≤ 0 are associated with stabilising mutations.
3. *Protocol 3:* We used the PRODIGY server[70] to calculate the binding energy for the human complex and for the 3D models of the other 215 animal complexes. We then calculated the change in binding energy from the human complex to the animal complex.

mCSM-PPI1 assigns a graph-based signature vector to each mutation, which is then used within machine learning models to predict the binding energy. The signature vector is based upon atom-distance patterns in the wild-type protein, pharmacophore information and available experimental information. The more recent mCSM-PPI2 is a development of the former with an improved description of the signature vector including, among other features, evolutionary information and energetic terms. For both mCSM-PPI1 and mCSM-PPI2 we used the mCSM server for the simulations.

As a third method we used the PRODIGY web-server which calculates the binding energy as a function of inter-residue contacts and the non-interacting surface of the residues. As this method delivers an

absolute binding energy, we calculate the relative binding energy as $\Delta\Delta G = \Delta G_{human} - \Delta G_{animal}$. The calculations, using the PRODIGY web-server, were performed at 298 and 310 Kelvin.

By exploring multiple protocols and tools we hoped to gain some insights into the dependence of our analyses, and insights derived from them, on the methodology used.

Change in residue chemistry for mutations

We computed the Grantham score[71], to measure the degree of chemical change associated with mutations occurring in the key interface residues, for each vertebrate compared to the human sequence. For the ACE2-TMPRSS2 comparison we used the same set of organisms as for ACE2, but only those that had at least one TMPRSS2 protein available from ENSEMBL, thus reducing the number of available organisms to 156.

The sums and averages of Grantham scores were obtained for the contact residues and extended residues set in ACE2, and for the active site, cleavage site and shells at 5Å and 8Å from the active site for TMPRSS2. Data and plots for the various combinations are available in Supplementary Table S3.

Phylogeny of SARS-like betacoronaviruses

Genome assemblies were downloaded for a subset of SARS-CoV (n = 10), SARS-like (n = 27) and SARS-CoV-2 (n = 13) viruses from publicly available data on NCBI[72–79] and GISAID[15,16]. For a full list of metadata, as compiled by NextStrain[80], including contributing and submitting laboratories, see Supplementary Table S4. All assemblies were downloaded and aligned against the SARS-CoV-2 reference genome Wuhan-Hu-1 (NCBI NC_045512.2) using MAFFT[81]. The alignment was manually inspected and the first 130 bp and last 50 bp were masked. A maximum likelihood phylogenetic tree was built on the alignment using RaXML-NG[82] under the GTR+G substitution model with 100 bootstraps to assess branch support. The resulting phylogeny was plotted using ggtree[83] v1.16.6. The spike ORF (21563:25384 nt relative to Wuhan-Hu-1) was extracted and translated to amino acid sequence using Ape[84] v5.3. The alignment of spike protein sequences, and receptor binding domain, were assessed and visualised using ggtree.

Results

Conservation of ACE2 in vertebrates

We aligned protein sequences of 247 vertebrate orthologues of ACE2. Most orthologues have more than 60% sequence identity with human ACE2 (Figure 1A). For each orthologue, we generated a 3D model of the protein structure from its protein sequence using FunMod[67,68]. We were able to build high-quality models for 236 vertebrate orthologues, with nDOPE scores > -1 (Supplementary Table S5). 11 low-quality models were removed from the analysis. After this, we removed 21 models from the analysis that were missing > 10 DCEX residues, leaving 215 models to take forward for further analysis. The structural models were compared to the human ACE2 protein structure using SSAP[85]. Most models had high structural similarity to human ACE2 (Figure 1B and Figure 1C). However, some models, despite sharing > 60% sequence identity to the human ACE2, had low structural similarity (Figure 1C). These changes probably reflect mutations slightly altering the relative orientations of the S-protein and ACE2 across the interface. (Supplementary Figure S2).

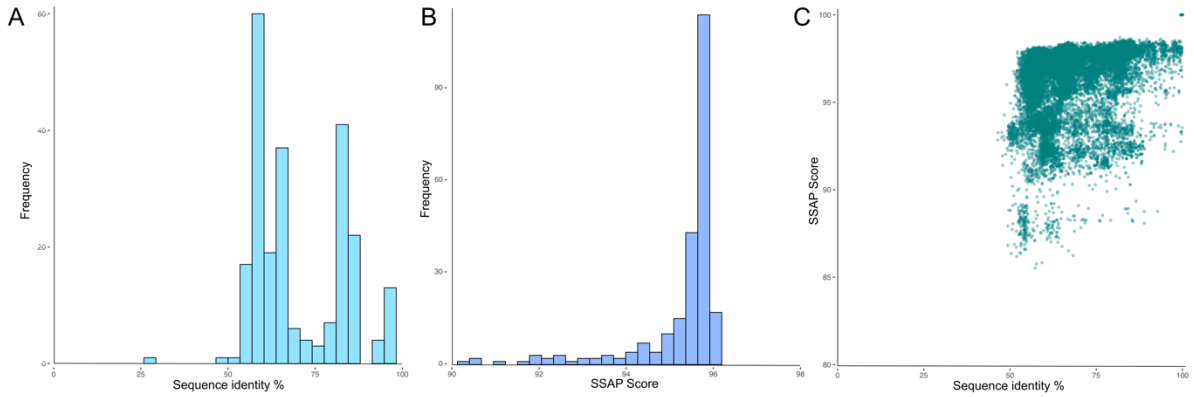

Figure 1. (a) Distribution of pairwise sequence identities of each vertebrate compared against the human sequence. (b) Distribution of SSAP scores of each vertebrate compared against the human sequence. (c) Pairwise sequence identities and SSAP scores for each organism in the dataset.

Identification of critical S-protein:ACE2 interface residues

Residues in the binding interface between the S-protein and ACE2 were identified in a structure of the complex (PDB ID 6M0J; Figure 2a, Supplementary Figure S1). Residues within 5Å of the interacting protein are reported in PDBe and PDBsum[55,86], which we refer to as 'Direct Contact' (DC) residues. Other important residues likely to influence binding have been identified using a variety of approaches, including structural analysis by us and other groups, alanine scanning, and mutagenesis[20,28,36,44,54,57,87]. In addition we identified highly conserved residues, sites under positive selection and predicted allosteric sites. This larger set includes 42 residues that are not involved in direct contact with the interacting protein but could still influence the binding affinity. 15 of these residues are within 5Å and 27 residues are further away but within 8Å, of one, or more, DC residues. We therefore compiled a set of residues that includes these residues, together with the DC residues. We refer to this as the direct contact plus 'extended set' of residues, which we refer to as 'DCEX' residues (Figure 2b). Some of the ACE2 DCEX residues are highly-conserved positions in the ACE2 FunFam (see Supplementary Figure S1 for summary of evidence for each residue in the DCEX set).

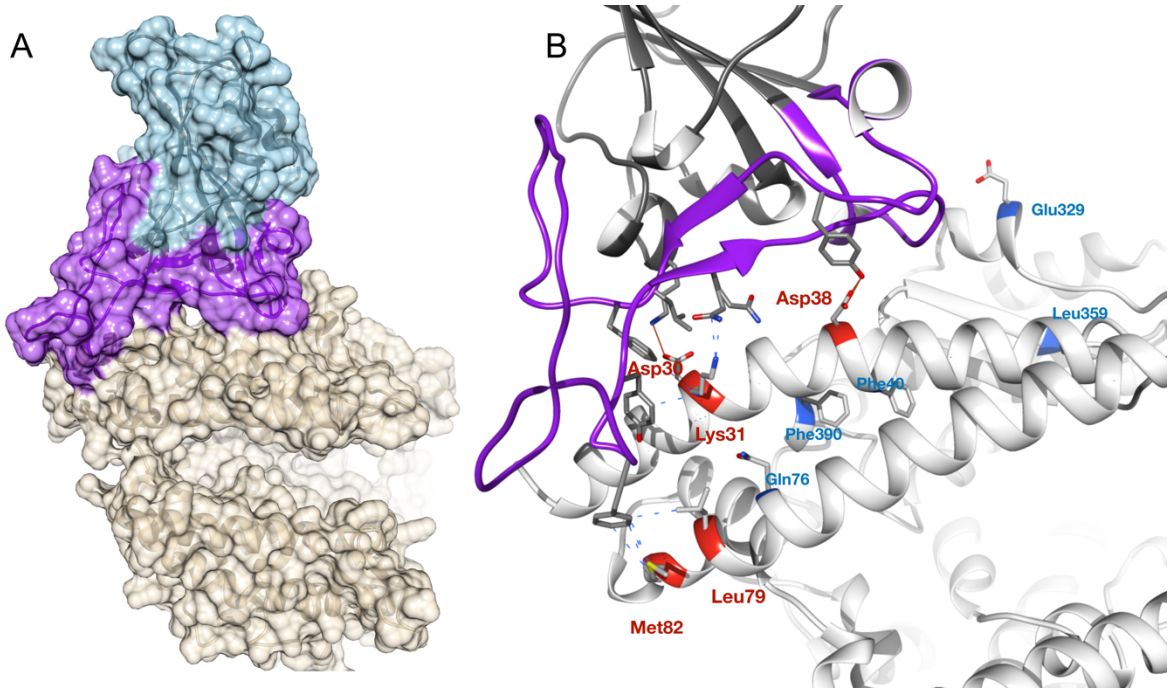

Figure 2. Overview of SARS-CoV-2 S-protein:human ACE2 complex and interface. (a) Interface of SARS-CoV-2 S-protein (light blue). The S-protein RBD (purple) binds to ACE2 (tan). (b) DC residues in the SARS-CoV-2 S-protein:human ACE2 complex interface. A subset of contact residues that are not conserved in vertebrates are highlighted for DC (red) and DCEX residues (blue) (PDB ID: 6M0J).

Changes in the energy of the S-protein:ACE2 complex in vertebrates

We used various protocols (protocols 1 and 2) to assess the relative change in binding energy (ΔΔG) of the SARS-CoV-2 S-protein:ACE2 complex following mutations in the DC and DCEX residues. Since these residues lie directly in the binding interface (DC residues), or in the secondary shell and are likely to be influencing binding, we hypothesise that the relative ΔΔG values obtained for mutations from animal sequence to human, can be used to gauge the relative risk of infection, for each animal. Our third protocol directly measured the binding energies for SARS-CoV-2 S-protein:ACE2 complex for the human structure and the modelled structures of each animal.

We examined how well these approaches correlated with each other (Supplementary Figure S3-S4-S5). We also correlated the ΔΔG values (protocols 1 and 2) and binding energies (protocol 3) with SARS-CoV-2 infection phenotypes in different animals[2,17–21,88,89] (Supplementary Table S6); NB some phenotype data are from studies that have not been peer reviewed). We found that protocol 2 employing mCSM-PPI2 (referred to as P(2)-PPI2 henceforth), calculated over the DCEX residues, correlated best with the phenotype data (Supplementary Figure S6), justifying the use of animal models to calculate ΔΔG values in this context. Since this protocol considers mutations from animal to human, lower ΔΔG values correspond to stabilisation of the animal complex and higher risk of infection. To consider these relative ΔΔG values in an evolutionary context, we annotated phylogenetic trees for all the 215 vertebrate species analysed (Supplementary Figure S7). Figure 3 shows the phylogenetic tree for a subset of these 215 species that humans come into close contact with in domestic, agricultural or zoological settings (human contact) together with the ΔΔG values, the number of residue changes compared to human ACE2 sequence and the total chemical shift across for the DCEX interface residues, for each animal. We calculate whether DC and DCEX residues

that are mutated in these species stabilise or destabilise the SARS-CoV-2 S-protein:ACE2 complex (Supplementary Figure S8 and Supplementary Figure S9).

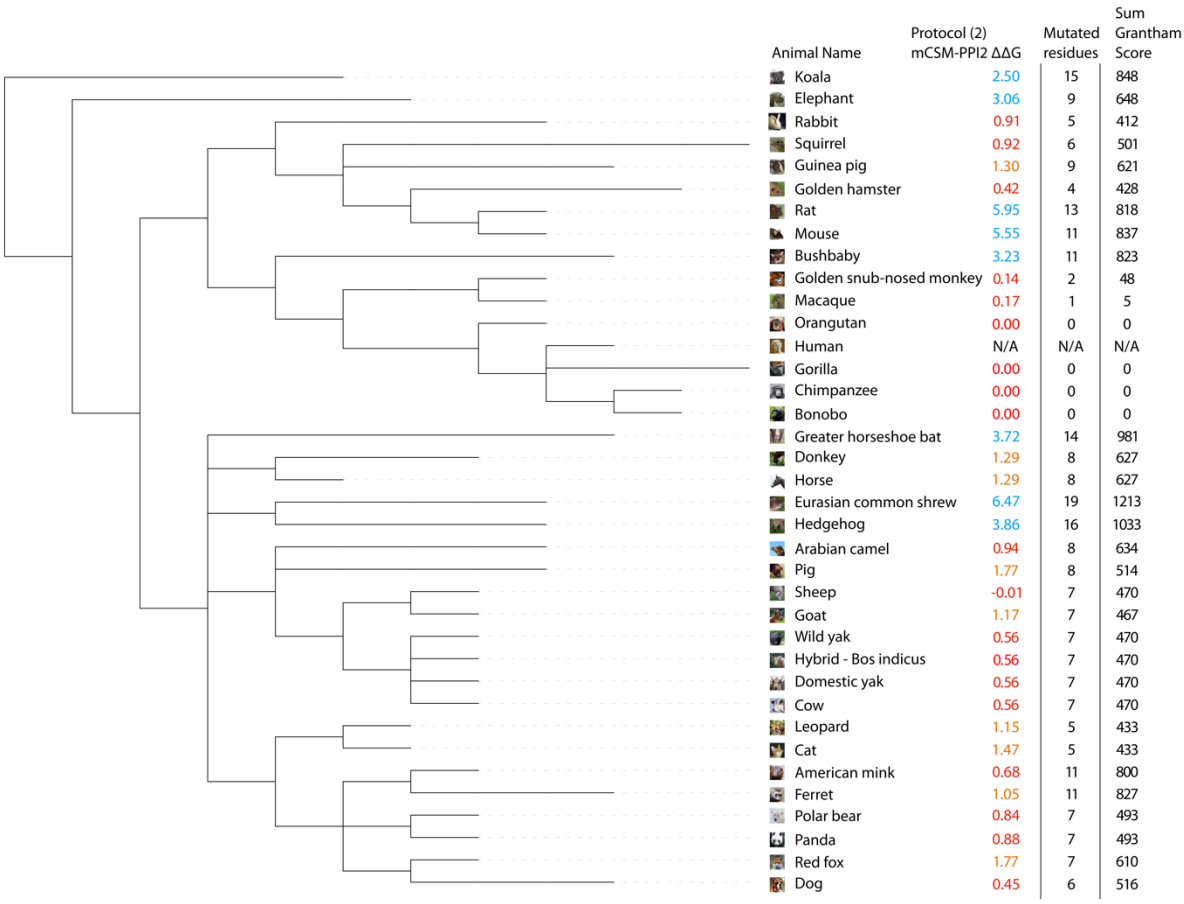

Figure 3 Phylogenetic tree of species that humans come into close contact with in domestic, agricultural or zoological settings. Leaves are annotated by the change in energy of the complex (ΔΔG), as measured by protocol 2 - PPI2. Animals are categorised according to risk of infection by SARS-CoV-2, with ΔΔG ≤ 1 high risk (red), 1 < ΔΔG < 2 medium risk (orange) and ΔΔG ≥ 2 low risk (blue). This tree contains a subset of animals from the larger ENSEMBL tree (Supplementary Figure S7).

The ΔΔG values measured by P(2)-PPI2 correlate well with the infection phenotypes (Supplementary Table S6). As shown in previous studies, many primates are at high risk[18,43,44], with a few exceptions including New World monkeys, for which the capuchin and the squirrel monkey all show no infection risk in experimental studies, in agreement with our predicted energies for the complex[19]. Contrary to our findings, an *in vivo* study showed that marmosets can be infected[18].

According to our relative ΔΔG values, zoological animals that come into contact with humans, such as pandas, leopards and bears, are also at risk of infection. In the agricultural context, camels, cows, sheep, goats and horses also have relatively low ΔΔG values suggesting comparable binding affinities to humans, as validated by the experimental data[19] (Supplementary Table S6). Whilst in domestic settings, dogs, cats, hamsters, and rabbits are also at risk from infection.

In general, we see a high infection risk for most mammals, with a notable exception for all non-placental mammals. Importantly, mice and rats are not susceptible, so hamsters and ferrets are being used as models of human COVID-19, instead of mice.

Of the 35 birds tested only the blue tit shows an infection risk. Similarly, the Nile tilapia is the only fish out of the 72 in this study which has a binding energy favorable to infection. Also, all 14 reptiles and amphibians we investigated do not show any risk.

The analysis shows that, with few exceptions, placental mammals are most susceptible to infection by SARS-CoV-2. Our predictions do not always agree with the experimental data. For example, there are discrepancies for guinea pig and donkey, where we predict moderate risk but no *in vitro* interaction was observed, and horseshoe bat, which we consider low risk but has *in vivo* evidence for infection. We consider these further below using detailed structural analyses.

Structural analyses of S-protein:ACE2 interactions

Cross-species comparison

We mapped DC residues and a selection of residues in our DCEX set that are within 8Å to the S-protein:ACE2 interface onto a 3D structure of the complex (PDB ID 6M0J[28]). The S-protein RBD contains a receptor binding motif (RBM), and five DC residues (positions 30, 31, 38, 79 and 82) and five DCEX residues (40, 76, 329, 359 and 390) that are not conserved between species (Figure 4a, Supplementary Figure S9).

Structural variations in the S-protein:ACE2 interface amongst models of eight species (macaque, guinea pig, camel, cat, cow, dog, sheep and horseshoe bat) were highlighted with reference to the human PDB structure (PDB ID 6M0J) and in the context of three regions of the interface: hydrophobic pocket, hotspot-353 and hotspot-31[29] (Figure 4a); these sites have previously been identified[29] as key to understanding why the SARS-CoV-2 S-protein binds to human ACE2 with high affinity and how the viral S-protein has evolved to bind with much higher affinity to human ACE2 than SARS-CoV[28,29,33]. Additional analysis of a putative allosteric site is discussed below. These animals were selected either because as they are likely to come into frequent contact with humans in domestic or agricultural setting (cat, dog, macaque, cow, sheep and camel) or because there were discrepancies between our energy calculations and the experimental data to date (guinea pig and horseshoe bat).

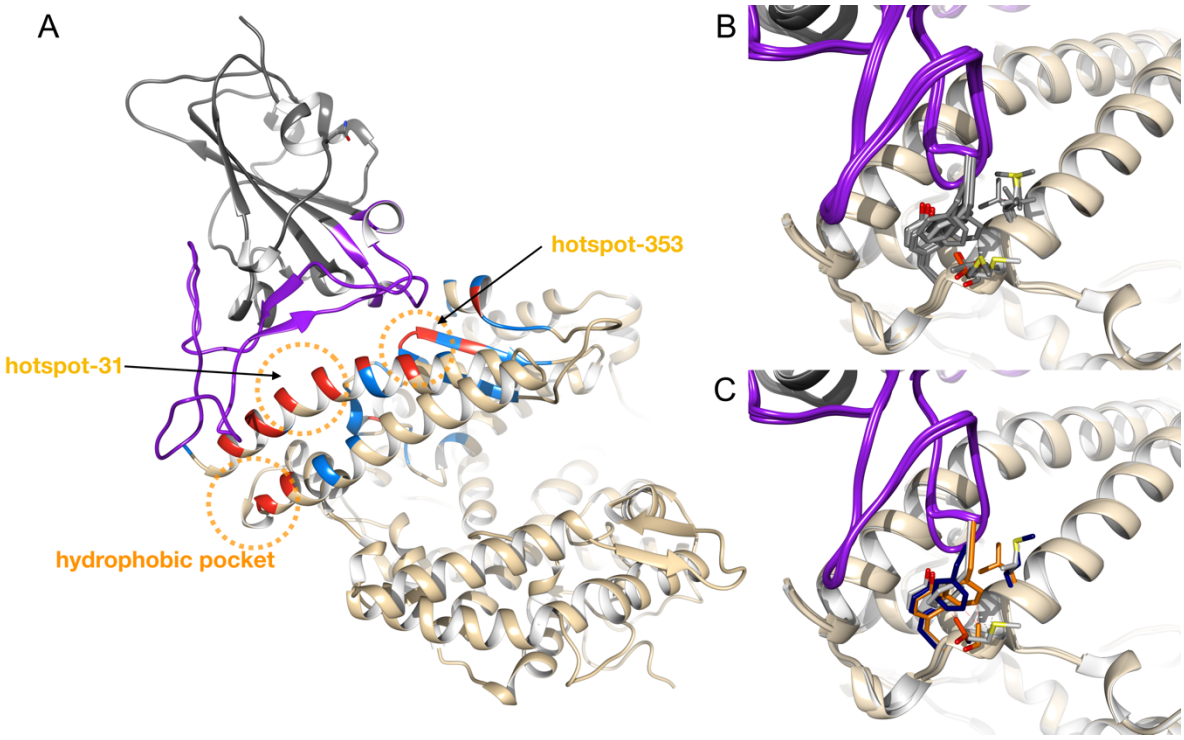

Figure 4: Structural features of the S-protein from SARS-CoV-2 and SARS-CoV bound to ACE2. (a) Three key interaction sites for the SARS-CoV-2 and SARS-CoV S-proteins with ACE2. (b) Variations in hydrophobic pocket residues at the S-protein:ACE2 interface and the range of conformations adopted by RBD residue Phe486 in 7 species. (c) Clash observed between RBD Phe486 and hydrophobic pocket in both dog and sheep due to non-conservation of pocket residues and changes in the conformation adopted by the RBD loop.

Hydrophobic pocket

The S-protein in SARS-CoV-2 has increased flexibility of the ACE2 interface loop due to a four residue motif (Gly-Val-Glu-Gly) in place of three in SARS-CoV (Pro-Pro-Ala), allowing for a more compact interface and insertion of SARS-CoV-2 RBD's Phe486 into a hydrophobic pocket in human ACE2 comprising Leu79, Met82 and Tyr83[29]. We assessed conformations of S-protein Phe486 and these three pocket residues for each of eight species compared to human. Phe486 binds in a similar conformation with respect to the hydrophobic pocket in human, macaque, guinea pig, camel, cat and cow (Figure 4b). However, both dog and sheep show a twisted orientation of the phenyl moiety, which clashes with Thr82 (Figure 4c) and would possibly reduce the binding affinity contribution of the pocket in these species.

Hotspot-353

ACE2 Lys353 is highly conserved and a key mediator of the SARS-CoV and SARS-CoV-2 S-protein interfaces[29]. Human Lys353 forms a salt bridge with Asp38 and these residues form H-bonds with the S-protein's RBD (Figure 5a). These interface interactions are significantly disrupted in horseshoe bat which has uncharged residue Asn38 in place of aspartate. This leads to loss of the salt bridge with Lys353 and a change in its conformation that no longer has a H-bond with S-protein Gly496 due to an increase in residue distance to 4.49Å (Figure 5b).

In cat and dog, the physicochemically similar but bulkier glutamate replaces aspartate, leading to loss of both the salt bridge and H-bond interactions at this position. However, we observed a reconfiguration of ACE2 residues permitting alternative H-bonds to form the S-protein RBD residues (Figure 5c).

Hotspot-31

In this hotpot, residues Lys31 and Glu35 are highly conserved. Variants at Lys31 are Glu31 (guinea pig) and Asp31 (horseshoe bat and camel). Residues on the S-protein in contact with position 31 are hydrophobic. Somewhat diverse conformations were observed at this site in guinea pig, horseshoe bat and camel (Supplementary Figure S10). In guinea pig Glu35 to Lys35 appears to no longer support H-bonds due to a pronounced change in conformation. Therefore, changes at this position are harder to interpret than for the other key sites.

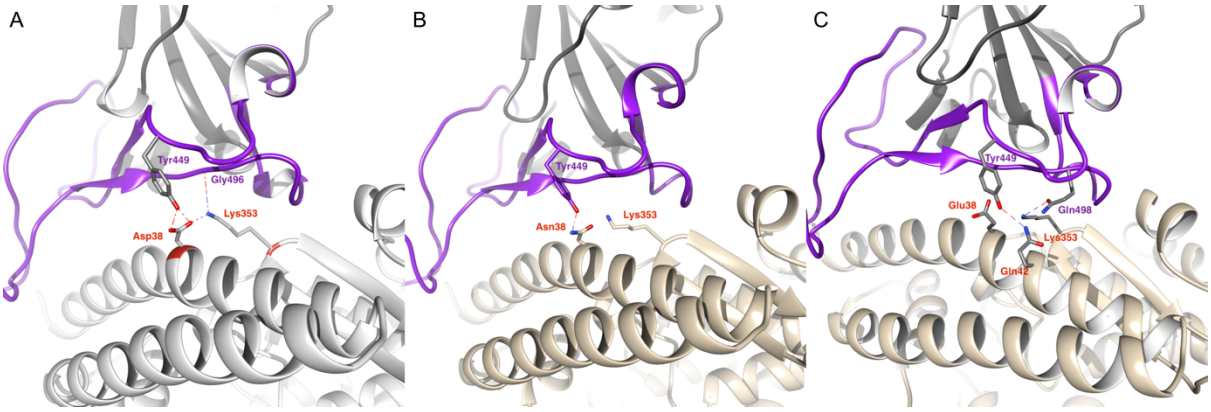

Figure 5: S-protein-ACE2 Interface interactions near the Lys353 hotspot are altered in species with Asp38 variants. (a) Human hotspot with salt-bridge Lys353-Asp38 and H-bonds to S-protein Tyr449 and Gly496. (b) Horseshoe bat Asn38 forms only H-bond with Tyr449. (c) In cat and dog bulkier Glu38 no longer forms salt bridge or H-bonds, but alternative interactions are predicted as H-bonds Lys353-Gln498 and Gln42-Tyr449.

Potential allosteric region formed by non-interface residues

We identified a cluster of residues in our DCEX set (Figure 6), in close proximity to the well-characterized hotspot residue K353 of ACE2. These residues lie in the β3 and β4 strands that connect the linker region containing K353. Most of these residues are highly conserved (ScoreCons scores > 0.70 and 10 residues (59%) have ScoreCons scores > 0.9). The proximity of the residues to the interface suggests a possible allosteric role for these very highly conserved residues. Furthermore, various methods for predicting allosteric sites in protein structures confirmed that this region is likely to be involved in allosteric regulation of the binding (see Supplementary Text for more details). Whilst most of these are highly conserved, for 168 of the animals there are 1 or more mutations in certain sites including 323-325, 329, 332, 333, 350, 352, 356 and 359, which could affect binding. More than 100 animals have mutations in three, or more, of these sites, however most of these changes are quite conservative. The only significant change is the L359K mutation, which moderately stabilises the complex in many animals.

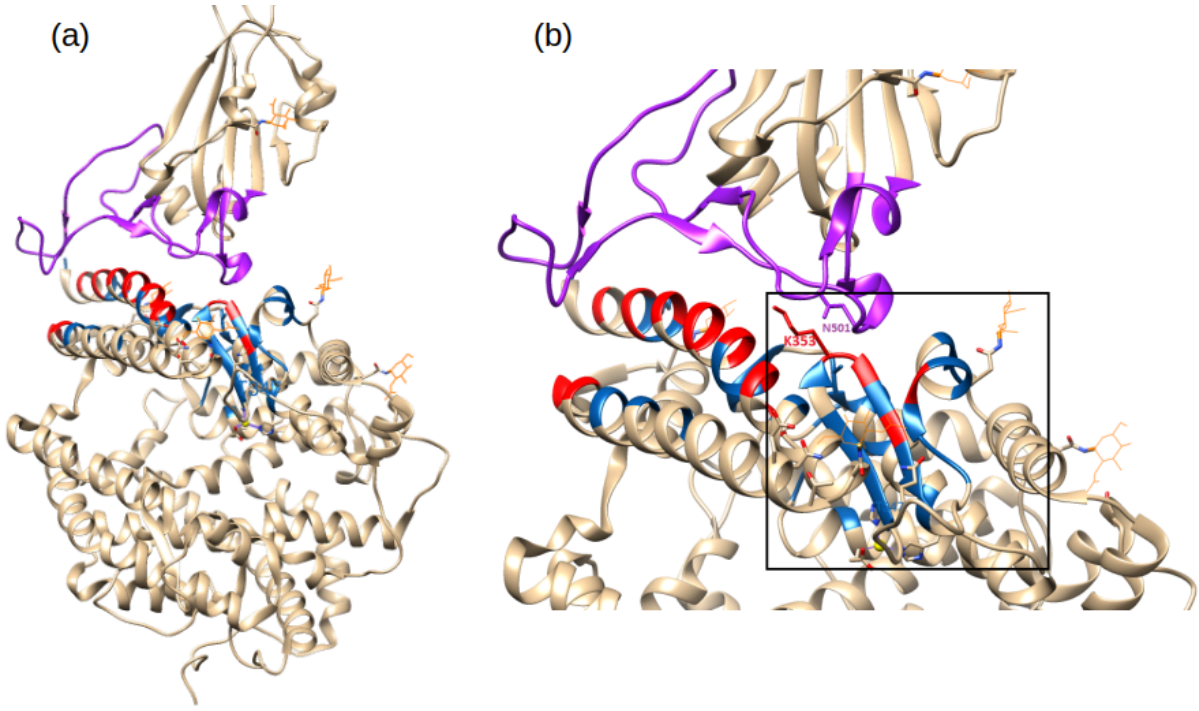

Figure 6: Potential allosteric region, identified in this study: The potential allosteric cluster (shown in blue) formed by non-interface residues is indicated in the box. The cluster is formed near the known hotspot residue K353 in ACE2. The cluster is formed by residues (β3: 347-350, 352), (β4: F356, I358, L359, M360), which connect to the linker region containing the hotspot residue K353. Additionally, residues, M323, T324, Q325, S331, M332, L333, W328 and E329 also form part of the cluster.

Changes in energy of the S-protein:ACE2 complex in SARS-CoV-2 and SARS-CoV

Farmed civets were infected by SARS-CoV in the 2002-2004 SARS epidemic[90], are thought to have been intermediate hosts in zoonosis[91], and thousands were culled in China to help to control the epidemic[92]. Using the SARS-CoV S-protein sequence, we analysed changes in energy of the S-protein:ACE2 complex for the 215 animals in our study. We used a structure of the SARS-CoV S-protein:ACE2 complex (PDB ID 2AJF) to model structures for the other animal species. Changes in energy of the S-protein:ACE2 complex are highly correlated in SARS-CoV-2 and SARS-CoV (Figure 7 and Supplementary Figure S11), suggesting that the range of animals susceptible to the virus is likely to be similar for SARS-CoV-2 and SARS-CoV.

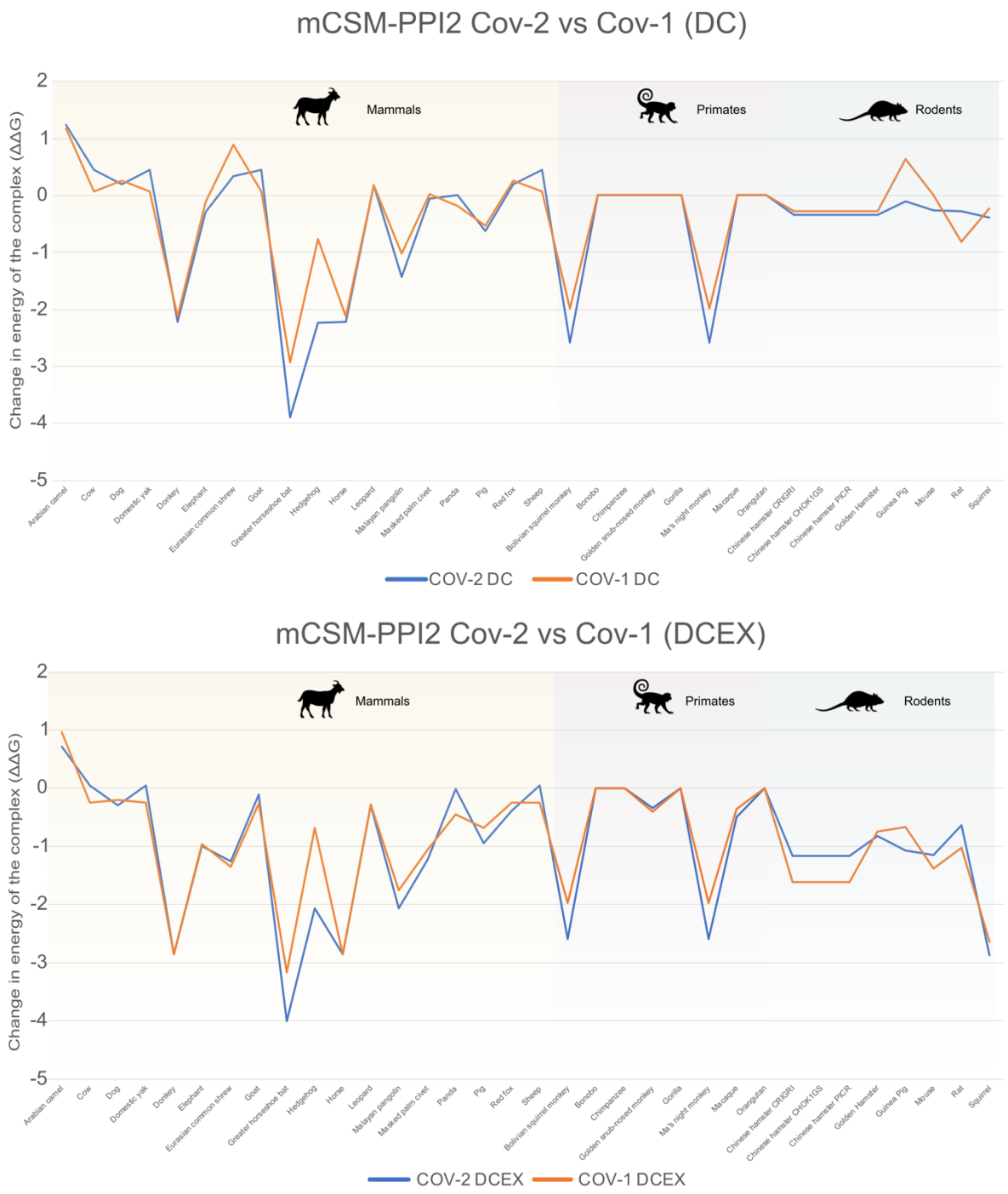

Figure 7: changes in the energy of the SARS-CoV:ACE2 and SARS-CoV-2:ACE2 complexes across selected animals. The change in energy of the complex (ΔΔG), as measured by protocol (2) - PPI2, is shown on the y-axis.

It should be noted that the binding of the SARS-CoV-2 S-protein:ACE2 complex is 10-22-fold stronger than for SARS-CoV[28,29,33], however both energies are sufficient for an interaction and clearly did enable infection with human hosts.

Conservation of TMPRSS2 and its role in SARS-CoV-2 infection

ACE2 and TMPRSS2 are key factors in the SARS-CoV-2 infection process. Both are highly co-expressed in the susceptible cell types, such as type II pneumocytes in the lungs, ileal absorptive enterocytes in the gut, and nasal goblet secretory cells[24]. Since both proteins are required for infection of host cells and since our analyses clearly support suggestions of conserved binding of S-protein:ACE2 across animal species, we decided to analyse whether the TMPRSS2 was similarly conserved. There is no known structure of TMPRSS2, but we built a high-quality model from a template structure (PDB ID 5I25). Since TMPRSS2 is a serine protease, and the key catalytic residues are known, this allowed us to use FunFams to identify highly conserved residues in the active site and the cleavage site that are likely to be involved in substrate binding. We therefore used two sets of residues for our analysis. The active site and cleavage site residues (ASCS) and the active site and cleavage site plus extended residues (ASCSEX) within 8Å of catalytic residues.

The average Grantham score for mutations of residues in the active site and cleavage site for TMPRSS2 is lower than ACE2 in all organisms under consideration (Figure 8). This means that the mutations TMPRSS2 involve more conservative changes. We expanded the analyses by including residues not further than 8Å and reached the same general result.

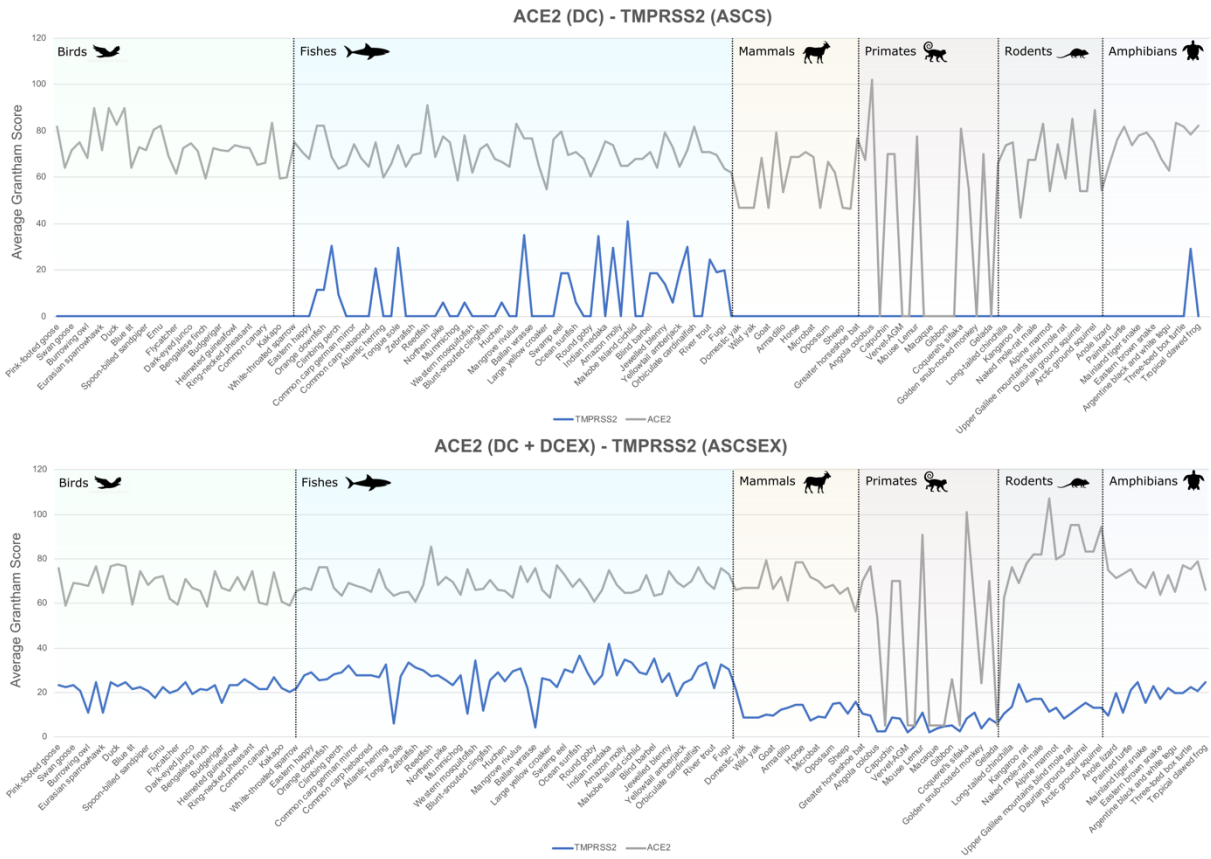

Figure 8: Comparison of Average Grantham Scores for ACE2 and TMPRSS2 for ASCS and ASCSEX residues

Mutations in ACE2 seem to have a more disruptive effect, and while we would expect a general conservation and lower Grantham score for organisms close to humans, there are still residue substitutions that have high Grantham scores > 70 for some primates, such as lemurs, tarsiers and

Bolivian squirrel monkey. These animals may therefore be less susceptible to infection because of these mutations in the protease.

Phylogenetic Analysis of SARS-like strains in different animal species

A small-scale phylogenetic analysis was performed on a subset of SARS-CoV-2 assemblies in conjunction with a broader range of SARS-like betacoronaviruses (Supplementary Table S3), including SARS-CoV isolated from humans and civets. The phylogeny is consistent with previous work[2] which identified the virus sampled from horseshoe bats (RaTG3, EPI_ISL_402131) as the closest genome sequence to SARS-CoV-2 strains in current global circulation (Figure 9). Aided by a large community effort, thousands of human-associated SARS-CoV-2 genome assemblies are now available on GISAID[15,16]. To date, these also include one assembly generated from a virus infecting a domestic dog (EPI_ISL_414518), one obtained from a zoo tiger (EPI_ISL_420923) and one obtained from a mink (EPI_ISL_431778). Each SARS-CoV-2 animal-associated strain falls within the phylogenetic diversity observed in human lineages (Figure 9a), providing evidence for recent human to animal transmission of SARS-CoV-2. The receptor binding domain is completely conserved (Figure 9b-c) in both human and animal SARS-CoV-2, with replacements in the spike protein of dog (S-protein V8L), tiger (S-protein D614G) and mink (S-protein D614G) strains relative to Wuhan-Hu-1 also observed in circulating human-associated lineages[93], again consistent with a recent transmission to non-human hosts. Of note, while genome-wide data indicates a closer phylogenetic relationship between SARS-CoV-2 strains and RatG13, the receptor binding domain alignment instead supports a closer relationship with a virus isolated from pangolins (EPI_ISL_410721; Figure 9c), in line with previous reports[94]. This highlights the importance of considering variation in putative structural determinants of host range, which may support a different topology to that observed based solely on genome-wide analyses.

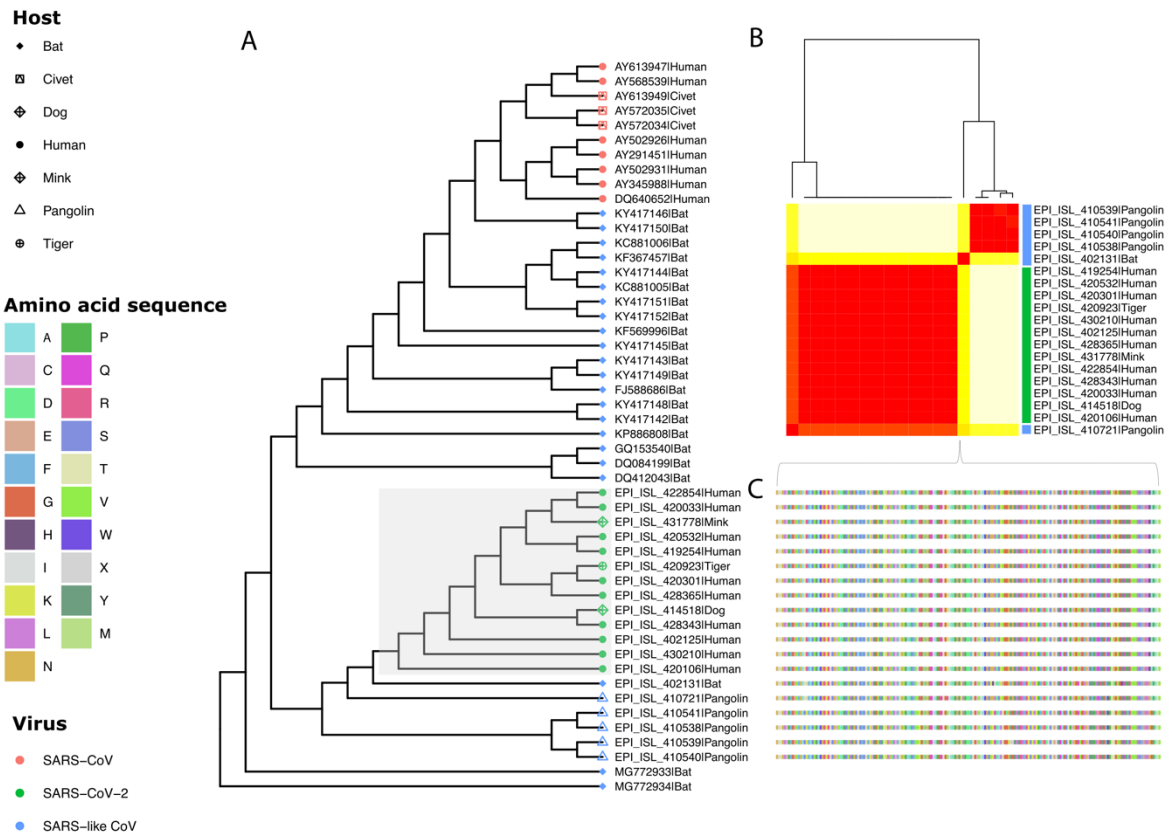

Figure 9: Phylogeny of SARS-like viruses. (a) Genome-wide maximum likelihood phylogenetic tree of SARS-like betacoronavirus strains sampled from diverse hosts (coloured tip symbols; Supplementary Table S3). NB genome EPI_ISL_402131 is from a horseshoe bat. (b) Pairwise amino acid differences at the S-protein RBD between human and animal strains of SARS-CoV-2, relative to closely related-SARS-like viruses in animal hosts. (c) Sequence alignment for the spike protein receptor binding domain.

Discussion

The ongoing COVID-19 global pandemic has a zoonotic origin, necessitating investigations into how SARS-CoV-2 infects animals, and how the virus can be transmitted across species. Whilst the binding energy of the complex formed between the S-protein and its receptors contributes to the viral host range, zoonosis and anthroponosis, so there is a clear need to study these interactions. However, relative changes in the energies of the S-protein:ACE2 complex have not been explored experimentally or *in silico*. A number of recent studies[19,20,43] have suggested that, due to high conservation of ACE2, some animals are vulnerable to infection by SARS-CoV-2. Concerningly, these animals could, in theory, serve as reservoirs of the virus, increasing the risk of future transmission across species. However, transmission rates across species are not known. Therefore, it is important to try to predict which other animals could potentially be infected by SARS-CoV-2, so that the plausible extent of zoonotic transmission can be estimated, and surveillance efforts can be guided.

Animal susceptibility to infection by SARS-CoV-2 has been studied *in vivo*[17,18,88,95,96] and *in vitro*[2,19,20] during the course of the pandemic. Parallel *in silico* work has made use of the protein structure of the S-protein:ACE2 complex to computationally predict the breadth of possible viral hosts. Most studies simply considered the number of residues mutated relative to human ACE2[32,97,98], although some also analyse the effect that these mutations have on the interface stability[35,44,99]. The most comprehensive of these studies analysed the number, and locations, of mutated residues in ACE2 orthologues from 410 species[43], but did not perform any structural analyses of the binding interface, except for identifying optimal rotamers of mutant residues using the human structure, and no assessment of TMPRSS2 was made. Furthermore, our work is the only study that has so far explored changes in the energy of the S-protein:ACE2 complex on a large scale.

In this study, we performed a comprehensive analysis of the major proteins that SARS-CoV-2 uses for cell entry. We predicted structures of ACE2 and TMPRSS2 orthologues from 215 vertebrate species and modelled S-protein:ACE2 complex and analysed changes in energy of the S-protein:ACE2 complex (ΔΔG), *in silico,* following mutations from animal residues to those in human. Our predictions suggest that, whilst many mammals are susceptible to infection by SARS-CoV-2, birds, fish and reptiles are not likely to be. We manually analysed residues in the S-protein:ACE2 interface, including DC residues that directly contacted the other protein, and DCEX residues that also included residues within 5Å of binding residues and other residues up to and including 8Å that may be important for binding. We identified these important residues by detailed manual inspection of the complex, evolutionary conservation patterns, *in silico* alanine scanning, allosteric predictions and detection of sites under positive selection. We clearly showed the advantage of performing more sophisticated studies of the changes in energy of the complex, over more simple measures (e.g. number and nature of mutated residues) used in other studies. Furthermore, the wider set of DCEX residues that we identified, near the binding interface, had a higher correlation to the phenotype data than the DC residues. In addition to ACE2, we also analysed how mutations in TMPRSS2 impact binding to the S-protein. We found that mutations in TMPRSS2 are less disruptive than mutations in ACE2, indicating that binding interactions in the S-protein:TMPRSS2 complex in different species will not be affected.

To increase our confidence in assessing changes in the energy of the complex, we developed multiple protocols using different, established methods. We correlated these stability measures with experimental infection phenotypes in the literature, from *in vivo* and *in vitro* studies of animals. Results from mCSM-PPI1 and mCSM-PPI2 had high correlation. Protocol 2 correlates best with the number of mutations, chemical changes induced by mutations, and infection phenotypes, so we chose to focus our analysis on this protocol.

Our measurement of changes in energy of the complex, together with high sequence and structural identity of ACE2 and TMPRSS2 orthologues, suggests that a broad range of mammals—but not fish, birds or reptiles—are susceptible to infection by SARS-CoV-2. Our predictions are supported by the findings of experimental studies assessing species that are at risk of infection by SARS-CoV-2[2,17–21,88,89]. Humans are likely to come into contact with 26 of these species in domestic, agricultural or zoological settings (Figure 10). Of particular concern are sheep, that have no change in energy of the S-protein:ACE2 complex, because they are farmed and come into close contact with humans. We provide phylogenetic evidence that SARS-CoV-2 recently transmitted from humans to tigers[14], dogs[15,16] and mink[15,16] (Figure 9). Our measurements of the change in energy of the complex for the

SARS-CoV S-protein were highly correlated with SARS-CoV-2, so our findings are also applicable to SARS-CoV.

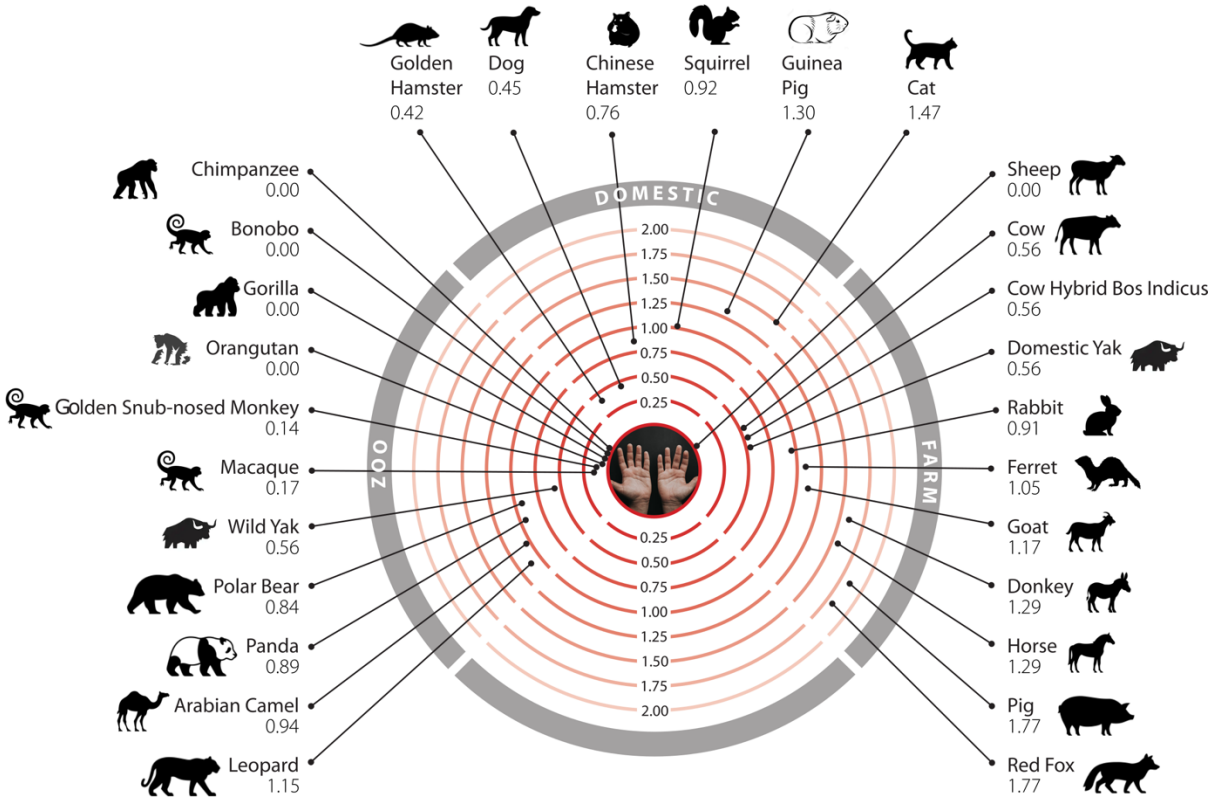

Figure 10 Mammals that humans come into contact with that are at risk of infection by SARS-CoV-2. 26 mammals are categorised into* domestic, agricultural or zoological settings*. Numbers represent the change in binding energy (ΔΔG) of the S-protein:ACE2.

For a subset of species, we performed more detailed structural analyses to gain a better understanding of the nature of the S-protein:ACE2 interface. In some cases, we had found discrepancies between our energy calculations and experimental phenotypes. To test our predictions, we manually analysed how the shape or chemistry of residues may impact complex stability for all DC residues and a selection of DCEX residues. In agreement with other studies[29], we identified a number of locations in ACE2 that are important for binding the S-protein. These locations, namely the hydrophobic cluster near residue the N-terminus, and two hotspot locations near residues 31 and 353, stabilise the binding interface. Five DC residues have species-specific variants and influence how well S-protein can bind utilising these key interface regions. In agreement with our calculations in changes in energy of the S-protein:ACE2 complex, our structural studies do not support *(in vivo)* infection of horseshoe bat, which has variants at three out of five variant DC residues, one of which (D38N) causes the loss of a salt bridge and H-bonding interactions between ACE2 and S-protein at hotspot 353. These detailed structural analyses are supported by the high Grantham score and calculated total ΔΔG for the change in energy of the complex. Both dog and cat have a physico-chemically similar variant at this hotspot (D38E), which although disrupting the salt bridge still permits alternative H-bonding interactions between the spike RBD and ACE2.

SARS-CoV-2 is better able to exploit the hydrophobic pocket than SARS-CoV by increased flexibility of its RBD loop and by mutation of L486F[29]. Our structures show how SARS-CoV-2 can utilise this pocket for binding at the interface in a wide range of species. Of species with DCEX variants at this pocket, only guinea pig maintains an entirely hydrophobic environment (M82A), whilst also conserving three out of the five variant DC residues. This helps explain why we predict moderate risk of infection, in contrast to the *in vitro* experimental data that reports no infection. For the marmoset, there is contradictory *in vivo* and *in vitro* experimental data. Our energy calculations suggest no risk and the structural analyses support this by identifying a large structural difference which could affect the complex. Finally, some DCEX residues were predicted to be allosteric sites, which may be promising drug targets[87].

The ability of SARS-CoV-2 to infect host cells to cause COVID-19, which sometimes results in severe disease, ultimately depends on a multitude of other host-virus protein interactions[100]. While we do not investigate them all in this study, our results suggest that SARS-CoV-2 can indeed infect a broad range of mammals. As there is a possibility of creating new reservoirs of the virus, we should now consider how to identify such transmission early and to mitigate against such risks. Animals living in close contact with humans should be monitored and farm animals should be protected where possible and managed accordingly (Santini & Edwards, under review).

Acknowledgments

We thank Gal Horesh and Caitlin Lee Carpenter for insightful discussions and Alan Hunns for help in making figures. We would also like to thank Francois Balloux, whose team member, Lucy van Dorp, contributed the phylogenetic analysis of SARS-like viruses.

Funding

HS is funded by Wellcome [203780/Z/16/A]. LvD acknowledges financial support from the Newton Fund UK-China NSFC initiative [MR/P007597/1] and a BBSRC equipment grant [BB/R01356X/1]. This work was also supported by other Wellcome (ND), BBSRC (NB, PA, NS, CR and IS) and UCL GCRF (VW) grants, codes are currently unavailable but will be added before submission.

Author contributions

SL, JL, JS, CO conceived the original idea. SL, CO, JL, LvD designed the experiments. SL, NB, VW, PA, LvD performed the experiments. SL, NB, VW, PA, NS, LvD, CR, IS, CO analysed data. SL, NB, VW, HS, PA, NS, LvD, CR, ND, IS, JL, CO interpreted the results. SL, NB, VW, HS, PA, NS, LvD, CR, ND, MP, MA, IS contributed to the manuscript and figures. HS, VW, PA, SL, CO wrote the manuscript. JL, SE, FF, JS, CO revised the manuscript.

References

1. World Health Organisation. COVID-19 situation report - 94.
2. Zhou, P. *et al.* A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature* **579**, 270–273 (2020).
3. Guan, Y. *et al.* Isolation and characterization of viruses related to the SARS coronavirus from animals in southern China. *Science* **302**, 276–278 (2003).
4. Wang, L. F. & Eaton, B. T. Bats, civets and the emergence of SARS. *Curr. Top. Microbiol. Immunol.* **315**, 325–344 (2007).
5. Shaw, L. P. *et al. The phylogenetic range of bacterial and viral pathogens of vertebrates*. http://biorxiv.org/lookup/doi/10.1101/670315 (2019) doi:10.1101/670315.
6. Saif, L. J. Bovine Respiratory Coronavirus. *Vet. Clin. North Am. Food Anim. Pract.* **26**, 349 (2010).
7. Vlasova, A. N. *et al.* Porcine Coronaviruses. *Emerg. Transbound. Anim. Viruses* 79–110 (2020) doi:10.1007/978-981-15-0402-0_4.
8. Lau, S. K. P. *et al.* Discovery of a Novel Coronavirus, China Rattus Coronavirus HKU24, from Norway Rats Supports the Murine Origin of Betacoronavirus 1 and Has Implications for the Ancestor of Betacoronavirus Lineage A. *J. Virol.* **89**, 3076–3092 (2015).
9. Lau, S. K. P. *et al.* Isolation and Characterization of a Novel Betacoronavirus Subgroup A Coronavirus, Rabbit Coronavirus HKU14, from Domestic Rabbits. *J. Virol.* **86**, 5481–5496 (2012).
10. Hasoksuz, M. *et al.* Biologic, antigenic, and full-length genomic characterization of a bovine-like coronavirus isolated from a giraffe. *J. Virol.* **81**, 4981–4990 (2007).
11. WHO | SARS (Severe Acute Respiratory Syndrome). *WHO* https://www.who.int/ith/diseases/sars/en/.
12. Chen, W. *et al.* SARS-associated Coronavirus Transmitted from Human to Pig. *Emerg. Infect. Dis.* **11**, 446–448 (2005).
13. Zhang, Q. SARS-CoV-2 neutralizing serum antibodies in cats: a serological investigation. https://www.biorxiv.org/content/10.1101/2020.04.01.021196v1.
14. USDA APHIS | USDA Statement on the Confirmation of COVID-19 in a Tiger in New York. https://www.aphis.usda.gov/aphis/newsroom/news/sa_by_date/sa-2020/ny-zoo-covid-19.
15. Elbe, S. & Buckland-Merrett, G. Data, disease and diplomacy: GISAID's innovative contribution to global health: Data, Disease and Diplomacy. *Glob. Chall.* **1**, 33–46 (2017).
16. Shu, Y. & McCauley, J. GISAID: Global initiative on sharing all influenza data - from vision to reality. *Euro Surveill. Bull. Eur. Sur Mal. Transm. Eur. Commun. Dis. Bull.* **22**, (2017).
17. Shi, J. *et al.* Susceptibility of ferrets, cats, dogs, and other domesticated animals to SARS–coronavirus 2. *Science* (2020) doi:10.1126/science.abb7015.
18. Lu, S. *et al.* Comparison of SARS-CoV-2 infections among 3 species of non-human primates. *bioRxiv* 2020.04.08.031807 (2020) doi:10.1101/2020.04.08.031807.
19. Liu, Y. *et al.* Functional and Genetic Analysis of Viral Receptor ACE2 Orthologs Reveals Broad Potential Host Range of SARS-CoV-2. *bioRxiv* 2020.04.22.046565 (2020) doi:10.1101/2020.04.22.046565.
20. Li, Y. *et al. Potential host range of multiple SARS-like coronaviruses and an improved ACE2-Fc variant that is potent against both SARS-CoV-2 and SARS-CoV-1*. http://biorxiv.org/lookup/doi/10.1101/2020.04.10.032342 (2020) doi:10.1101/2020.04.10.032342.

21. Zhao, X. *et al.* Broad and differential animal ACE2 receptor usage by SARS-CoV-2. *bioRxiv* 2020.04.19.048710 (2020) doi:10.1101/2020.04.19.048710.
22. Yu, P. *et al.* Age-related rhesus macaque models of COVID-19. *Anim. Models Exp. Med.* **3**, 93–97 (2020).
23. Li, W. *et al.* Angiotensin-converting enzyme 2 is a functional receptor for the SARS coronavirus. *Nature* **426**, 450–454 (2003).
24. Hamming, I. *et al.* Tissue distribution of ACE2 protein, the functional receptor for SARS coronavirus. A first step in understanding SARS pathogenesis. *J. Pathol.* **203**, 631–637 (2004).
25. Ziegler, C. *et al. SARS-CoV-2 Receptor ACE2 is an Interferon-Stimulated Gene in Human Airway Epithelial Cells and Is Enriched in Specific Cell Subsets Across Tissues*. https://papers.ssrn.com/abstract=3555145 (2020) doi:10.2139/ssrn.3555145.
26. Feng, Y. *et al.* ACE2 over-expression in the subfornical organ prevents the Angiotensin-II-mediated pressor and drinking responses and is associated with AT1 receptor down-regulation. *Circ. Res.* **102**, 729–736 (2008).
27. Patel, V. B., Zhong, J.-C., Grant, M. B. & Oudit, G. Y. Role of the ACE2/Angiotensin 1–7 axis of the Renin-Angiotensin System in Heart Failure. *Circ. Res.* **118**, 1313–1326 (2016).
28. Lan, J. *et al.* Structure of the SARS-CoV-2 spike receptor-binding domain bound to the ACE2 receptor. *Nature* 1–6 (2020) doi:10.1038/s41586-020-2180-5.
29. Shang, J. *et al.* Structural basis of receptor recognition by SARS-CoV-2. *Nature* (2020) doi:10.1038/s41586-020-2179-y.
30. Li, F., Li, W., Farzan, M. & Harrison, S. C. Structure of SARS Coronavirus Spike Receptor-Binding Domain Complexed with Receptor. *Science* **309**, 1864–1868 (2005).
31. Frank, H. K., Enard, D. & Boyd, S. D. Exceptional diversity and selection pressure on SARS-CoV and SARS-CoV-2 host receptor in bats compared to other mammals. *bioRxiv* 2020.04.20.051656 (2020) doi:10.1101/2020.04.20.051656.
32. Sun, J. *et al.* COVID-19: Epidemiology, Evolution, and Cross-Disciplinary Perspectives. *Trends Mol. Med.* **0**, (2020).
33. Wrapp, D. *et al.* Cryo-EM structure of the 2019-nCoV spike in the prefusion conformation. *Science* **367**, 1260–1263 (2020).
34. Ou, J. *et al. Emergence of RBD mutations in circulating SARS-CoV-2 strains enhancing the structural stability and human ACE2 receptor affinity of the spike protein*. http://biorxiv.org/lookup/doi/10.1101/2020.03.15.991844 (2020) doi:10.1101/2020.03.15.991844.
35. Stawiski, E. W. *et al. Human ACE2 receptor polymorphisms predict SARS-CoV-2 susceptibility*. http://biorxiv.org/lookup/doi/10.1101/2020.04.07.024752 (2020) doi:10.1101/2020.04.07.024752.
36. Hussain, M. *et al.* Structural Basis of SARS-CoV-2 Spike Protein Priming by TMPRSS2. *bioRxiv* 2020.04.21.052639 (2020) doi:10.1101/2020.04.21.052639.
37. Heurich, A. *et al.* TMPRSS2 and ADAM17 Cleave ACE2 Differentially and Only Proteolysis by TMPRSS2 Augments Entry Driven by the Severe Acute Respiratory Syndrome Coronavirus Spike Protein. *J. Virol.* **88**, 1293–1307 (2014).
38. Hoffmann, M. *et al.* SARS-CoV-2 Cell Entry Depends on ACE2 and TMPRSS2 and Is Blocked by a Clinically Proven Protease Inhibitor. *Cell* **181**, 271-280.e8 (2020).
39. Asselta, R., Paraboschi, E. M., Mantovani, A. & Duga, S. ACE2 and TMPRSS2 variants and expression as candidates to sex and country differences in COVID-19 severity in Italy. *medRxiv*

2020.03.30.20047878 (2020) doi:10.1101/2020.03.30.20047878.
40. Stopsack, K. H., Mucci, L. A., Antonarakis, E. S., Nelson, P. S. & Kantoff, P. W. TMPRSS2 and COVID-19: Serendipity or opportunity for intervention? *Cancer Discov.* (2020) doi:10.1158/2159-8290.CD-20-0451.
41. Jin, J.-M. *et al.* Gender differences in patients with COVID-19: Focus on severity and mortality. *medRxiv* 2020.02.23.20026864 (2020) doi:10.1101/2020.02.23.20026864.
42. Wenham, C., Smith, J. & Morgan, R. COVID-19: the gendered impacts of the outbreak. *The Lancet* **395**, 846–848 (2020).
43. Damas, J. *et al.* Broad Host Range of SARS-CoV-2 Predicted by Comparative and Structural Analysis of ACE2 in Vertebrates. *bioRxiv* 2020.04.16.045302 (2020) doi:10.1101/2020.04.16.045302.
44. Melin, A. D., Janiak, M. C., Marrone, F., Arora, P. S. & Higham, J. P. *Comparative ACE2 variation and primate COVID-19 risk*. http://biorxiv.org/lookup/doi/10.1101/2020.04.09.034967 (2020) doi:10.1101/2020.04.09.034967.
45. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).
46. Das, S. *et al.* Functional classification of CATH superfamilies: a domain-based approach for protein function annotation. *Bioinformatics* **31**, 3460–3467 (2015).
47. Lee, D. *et al.* Novel Computational Protocols for Functionally Classifying and Characterising Serine Beta-Lactamases. *PLoS Comput. Biol.* **12**, e1004926 (2016).
48. Ashford, P., Pang, C. S. M., Moya-García, A. A., Adeyelu, T. & Orengo, C. A. A CATH domain functional family based approach to identify putative cancer driver genes and driver mutations. *Sci. Rep.* **9**, 263 (2019).
49. Dawson, N. L. *et al.* CATH: an expanded resource to predict protein function through structure and sequence. *Nucleic Acids Res.* **45**, D289–D295 (2017).
50. Lewis, T. E., Sillitoe, I. & Lees, J. G. cath-resolve-hits: a new tool that resolves domain matches suspiciously quickly. *Bioinformatics* **35**, 1766–1767 (2019).
51. Higgins, D. G. & Sharp, P. M. CLUSTAL: a package for performing multiple sequence alignment on a microcomputer. *Gene* **73**, 237–244 (1988).
52. Waterhouse, A. M., Procter, J. B., Martin, D. M. A., Clamp, M. & Barton, G. J. Jalview Version 2--a multiple sequence alignment editor and analysis workbench. *Bioinforma. Oxf. Engl.* **25**, 1189–1191 (2009).
53. Valdar, W. S. J. Scoring residue conservation. *Proteins* **48**, 227–241 (2002).
54. Procko, E. The sequence of human ACE2 is suboptimal for binding the S spike protein of SARS coronavirus 2. *bioRxiv* 2020.03.16.994236 (2020) doi:10.1101/2020.03.16.994236.
55. Laskowski, R. A. PDBsum: summaries and analyses of PDB structures. *Nucleic Acids Res.* **29**, 221–222 (2001).
56. Pires, D. E. V., Ascher, D. B. & Blundell, T. L. mCSM: predicting the effects of mutations in proteins using graph-based signatures. *Bioinforma. Oxf. Engl.* **30**, 335–342 (2014).
57. Wan, Y., Shang, J., Graham, R., Baric, R. S. & Li, F. Receptor Recognition by the Novel Coronavirus from Wuhan: an Analysis Based on Decade-Long Structural Studies of SARS Coronavirus. *J. Virol.* **94**, e00127-20, /jvi/94/7/JVI.00127-20.atom (2020).
58. Pettersen, E. F. *et al.* UCSF Chimera--a visualization system for exploratory research and analysis. *J. Comput. Chem.* **25**, 1605–1612 (2004).
59. Luan, J., Lu, Y., Jin, X. & Zhang, L. Spike protein recognition of mammalian ACE2 predicts the

host range and an optimized ACE2 for SARS-CoV-2 infection. *Biochem. Biophys. Res. Commun.* **526**, 165–169 (2020).
60. Yan, R. *et al.* Structural basis for the recognition of SARS-CoV-2 by full-length human ACE2. *Science* **367**, 1444–1448 (2020).
61. Huang, W. *et al.* Allosite: a method for predicting allosteric sites. *Bioinformatics* **29**, 2357–2359 (2013).
62. Li, H., Chang, Y.-Y., Lee, J. Y., Bahar, I. & Yang, L.-W. DynOmics: dynamics of structural proteome and beyond. *Nucleic Acids Res.* **45**, W374–W380 (2017).
63. Panjkovich, A. & Daura, X. PARS: a web server for the prediction of Protein Allosteric and Regulatory Sites. *Bioinformatics* **30**, 1314–1315 (2014).
64. Song, K. *et al.* Improved Method for the Identification and Validation of Allosteric Sites. *J. Chem. Inf. Model.* **57**, 2358–2363 (2017).
65. Murrell, B. *et al.* Detecting Individual Sites Subject to Episodic Diversifying Selection. *PLOS Genet.* **8**, e1002764 (2012).
66. Weaver, S. *et al.* Datamonkey 2.0: A Modern Web Application for Characterizing Selective and Other Evolutionary Processes. *Mol. Biol. Evol.* **35**, 773–777 (2018).
67. Lam, S. D. *et al.* Gene3D: expanding the utility of domain assignments. *Nucleic Acids Res.* **44**, D404–D409 (2016).
68. Lam, S. D., Das, S., Sillitoe, I. & Orengo, C. An overview of comparative modelling and resources dedicated to large-scale modelling of genome sequences. *Acta Crystallogr. Sect. D* **73**, 628–640 (2017).
69. Rodrigues, C. H. M., Myung, Y., Pires, D. E. V. & Ascher, D. B. mCSM-PPI2: predicting the effects of mutations on protein–protein interactions. *Nucleic Acids Res.* **47**, W338–W344 (2019).
70. Xue, L. C., Rodrigues, J. P., Kastritis, P. L., Bonvin, A. M. & Vangone, A. PRODIGY: a web server for predicting the binding affinity of protein–protein complexes. *Bioinformatics* btw514 (2016) doi:10.1093/bioinformatics/btw514.
71. Grantham, R. Amino Acid Difference Formula to Help Explain Protein Evolution. *Science* **185**, 862–864 (1974).
72. Yeh, S.-H. *et al.* Characterization of severe acute respiratory syndrome coronavirus genomes in Taiwan: Molecular epidemiology and genome evolution. *Proc. Natl. Acad. Sci.* **101**, 2542–2547 (2004).
73. Chim, S. S. C. *et al.* Genomic characterisation of the severe acute respiratory syndrome coronavirus of Amoy Gardens outbreak in Hong Kong. *Lancet Lond. Engl.* **362**, 1807–1808 (2003).
74. 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).
75. Ge, X.-Y. *et al.* Isolation and characterization of a bat SARS-like coronavirus that uses the ACE2 receptor. *Nature* **503**, 535–538 (2013).
76. Li, W. *et al.* Bats are natural reservoirs of SARS-like coronaviruses. *Science* **310**, 676–679 (2005).
77. Lau, S. K. P. *et al.* Ecoepidemiology and Complete Genome Comparison of Different Strains of Severe Acute Respiratory Syndrome-Related Rhinolophus Bat Coronavirus in China Reveal Bats as a Reservoir for Acute, Self-Limiting Infection That Allows Recombination Events. *J. Virol.* **84**, 2808–2819 (2010).
78. Lau, S. K. P. *et al.* Severe acute respiratory syndrome coronavirus-like virus in Chinese horseshoe bats. *Proc. Natl. Acad. Sci. U. S. A.* **102**, 14040–14045 (2005).

79. Hu, D. *et al.* Genomic characterization and infectivity of a novel SARS-like coronavirus in Chinese bats. *Emerg. Microbes Infect.* **7**, 154 (2018).
80. Hadfield, J. *et al.* Nextstrain: real-time tracking of pathogen evolution. *Bioinforma. Oxf. Engl.* **34**, 4121–4123 (2018).
81. Nakamura, T., Yamada, K. D., Tomii, K. & Katoh, K. Parallelization of MAFFT for large-scale multiple sequence alignments. *Bioinforma. Oxf. Engl.* **34**, 2490–2492 (2018).
82. Kozlov, A. M., Darriba, D., Flouri, T., Morel, B. & Stamatakis, A. RAxML-NG: a fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference. *Bioinformatics* **35**, 4453–4455 (2019).
83. Yu, G., Smith, D. K., Zhu, H., Guan, Y. & Lam, T. T. ggtree:an R package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods Ecol. Evol.* **8**, 28–36 (2017).
84. Paradis, E. & Schliep, K. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinforma. Oxf. Engl.* **35**, 526–528 (2019).
85. Orengo, C. A. & Taylor, W. R. SSAP: sequential structure alignment program for protein structure comparison. *Methods Enzymol.* **266**, 617–635 (1996).
86. Armstrong, D. R. *et al.* PDBe: improved findability of macromolecular structure data in the PDB. *Nucleic Acids Res.* gkz990 (2019) doi:10.1093/nar/gkz990.
87. Han, Y. & Král, P. Computational Design of ACE2-Based Peptide Inhibitors of SARS-CoV-2. *ACS Nano* acsnano.0c02857 (2020) doi:10.1021/acsnano.0c02857.
88. Kim, Y.-I. *et al.* Infection and Rapid Transmission of SARS-CoV-2 in Ferrets. *Cell Host Microbe* S1931312820301876 (2020) doi:10.1016/j.chom.2020.03.023.
89. Rockx, B. *et al.* Comparative pathogenesis of COVID-19, MERS, and SARS in a nonhuman primate model. *Science* eabb7314 (2020) doi:10.1126/science.abb7314.
90. Tu, C. *et al.* Antibodies to SARS Coronavirus in Civets. *Emerg. Infect. Dis.* **10**, 2244–2248 (2004).
91. 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).
92. Watts, J. China culls wild animals to prevent new SARS threat. *The Lancet* **363**, 134 (2004).
93. CVR Bioinformatics. CoV-GLUE: Virus genome analysis for the SARS-CoV-2 pandemic. http://cov-glue.cvr.gla.ac.uk/#/home.
94. Zhang, T., Wu, Q. & Zhang, Z. Probable Pangolin Origin of SARS-CoV-2 Associated with the COVID-19 Outbreak. *Curr. Biol.* **30**, 1346-1351.e2 (2020).
95. Temmam, S. *et al. Absence of SARS-CoV-2 infection in cats and dogs in close contact with a cluster of COVID-19 patients in a veterinary campus*. http://biorxiv.org/lookup/doi/10.1101/2020.04.07.029090 (2020) doi:10.1101/2020.04.07.029090.
96. Shan, C. *et al. Infection with Novel Coronavirus (SARS-CoV-2) Causes Pneumonia in the Rhesus Macaques*. https://www.researchsquare.com/article/rs-15756/v1 (2020) doi:10.21203/rs.2.25200/v1.
97. Liu, Z. *et al.* 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 (2020).
98. Luan, J., Jin, X., Lu, Y. & Zhang, L. SARS-CoV-2 spike protein favors ACE2 from Bovidae and Cricetidae. *J. Med. Virol.* **n/a**,.
99. Jia, Y. *et al. Analysis of the mutation dynamics of SARS-CoV-2 reveals the spread history and*

emergence of RBD mutant with lower ACE2 binding affinity. http://biorxiv.org/lookup/doi/10.1101/2020.04.09.034942 (2020) doi:10.1101/2020.04.09.034942.

100. Gordon, D. E. *et al.* A SARS-CoV-2-Human Protein-Protein Interaction Map Reveals Drug Targets and Potential Drug-Repurposing. *bioRxiv* 2020.03.22.002386 (2020) doi:10.1101/2020.03.22.002386.