Journal Pre-proof

Architects of infection: A structural overview of SARS-related coronavirus spike glycoproteins

Francesca R. Hills, Jemma L. Geoghegan, Mihnea Bostina

PII: S0042-6822(24)00407-0

DOI: https://doi.org/10.1016/j.virol.2024.110383

Reference: YVIRO 110383

To appear in: *Virology*

Received Date: 20 October 2024

Revised Date: 22 December 2024

Accepted Date: 29 December 2024

Please cite this article as: Hills, F.R, Geoghegan, J.L, Bostina, M., Architects of infection: A structural overview of SARS-related coronavirus spike glycoproteins, *Virology*, https://doi.org/10.1016/j.virol.2024.110383.

Architects of infection: A structural overview of SARS-related coronavirus spike glycoproteins

Francesca R Hills[1], Jemma L Geoghegan[1], Mihnea Bostina[1]*
[1]Department of Microbiology and Immunology, University of Otago, Dunedin, New Zealand

**Corresponding author: Mihnea Bostina, 720 Cumberland Street, 9016, New Zealand*
mihnea.bostina@otago.ac.nz

Abstract

The frequency of zoonotic viral emergence within the *Coronaviridae* family highlights the critical need to understand the structural features of spike proteins that govern viral entry and host adaptation. Investigating the structural conservation and variation in key regions of the spike protein—those involved in host range, binding affinity, viral entry, and immune evasion—is essential for predicting the evolutionary pathways of coronaviruses, assessing the risk of future host-jumping events, and discovering pan-neutralising antibodies. Here we summarise our current structural understanding of the spike proteins similar to SARS-CoV-2 from the *Coronaviridae* family and compare key functional similarities and differences. Our aim is to demonstrate the significant structural and sequence conservation between spike proteins from a range of host species and to outline the importance of animal coronavirus surveillance and structural investigation in our endeavour for pandemic preparedness against emerging viruses.

Key words

Coronaviruses, spike protein, SARS-related coronaviruses, cryo-electron microscopy, structural biology

Introduction

With international travel, climate change, and increased human contact with wild animals, the ability for emerging zoonotic viruses to spread quickly through the human population has greatly increased in the last 50 years (Allen et al., 2017; Rahman et al., 2020). Zoonotic viral

emergence events have increased in frequency in recent years, with viruses from the *Coronaviridae* – SARS-CoV, MERS-CoV and SARS-CoV-2 – being responsible for three pandemics in the last two decades (Allen et al., 2017; Drosten et al., 2003; Kuiken et al., 2003; Rahman et al., 2020; Stadler et al., 2003; Zaki et al., 2012; Zhou et al., 2020). SARS-CoV and SARS-CoV-2 belong to the subgenus Sarbecovirus, of which horseshoe bats (Rhinolophus sp.) are the key reservoir species (Holmes et al., 2021; Wrobel et al., 2020; Ye et al., 2020; Zhou et al., 2020). Whilst a plethora of viral, host, and environmental factors play a role in the capability of a virus to infect both the reservoir and novel host species, the spike (S) glycoprotein specifically binds angiotensin-converting enzyme 2 (ACE2) and allows viral entry into a host cell (Hoffmann et al., 2020; Ou et al., 2020; Zhou et al., 2020). It is responsible for determining the viral host range and is a key antigenic site for the immune response (Li et al., 2003; Liu et al., 2021; Mittal et al., 2022; Xu et al., 2021; Zhou et al., 2020). Consequently, the structural investigation of S proteins has been mainly focused on bat coronaviruses (bCoV) (Fan et al., 2019; Lee et al., 2023; Menachery et al., 2015; Ou et al., 2023; Qiao and Wang, 2024; Wrobel et al., 2020; Xiong et al., 2022; Zhang et al., 2021c). Although SARS-CoV and SARS-CoV-2 are thought to originate in bats, it is likely these viruses reached the human population through intermediate hosts. For example, human ACE2 (hACE2) binding of bCoV_RaTG13, a close genetic relative to SARS-CoV-2, is poor and suggests there may have been recombination events within an intermediate host leading to the adaptation of effective hACE2 binding (Lam et al., 2020; Wrobel et al., 2020; Zhang et al., 2021c). While many potential intermediate hosts have been hypothesised (Crits-Christoph et al., 2024), to date only S proteins from two pangolin (pCoV) and two civet coronaviruses (cCoV) have been structurally investigated (Hills et al., 2024; Lam et al., 2020; Wrobel et al., 2021; Zhang et al., 2021c; Zhao et al., 2020).

The S protein follows the typical structural framework of a type I viral fusion protein with two main subunits, S1 and S2, both containing key structural domains contributing to virus-host binding, fusion, protein stability and antibody escape (Bosch et al., 2003; Walls et al., 2020; Wrapp et al., 2020). The S1 subunit houses the receptor binding domain (RBD) which includes a 70 amino acid-long receptor binding motif (RBM) that makes direct contact with the host receptor and the N-terminal domain (NTD) (Shang et al., 2020) (Figure 1a). Following the S1 subunit is the S1-S2 cleavage site, an essential motif for the series of conformational changes leading to the S protein post-fusion state following cleavage by host proteases (Belouzard et al., 2009; Lavie et al., 2022). The S2 subunit, which is highly conserved among

sarbecoviruses contains the fusion peptide (FP), heptad repeat 1 and 2 (HR1, HR2), the transmembrane domain (TM), and cytoplasmic domain (CP) (Wrapp et al., 2020) (Figure 1a).

This study reviews the currently available structures of spike glycoproteins from the Sarbecovirus subgenus and unclassified *Coronavirinae* (Table 1; Supplementary Table 2). Technical information, PDB accession codes and research articles related to Spike glycoprotein structures reviewed can be found in Table 1, with extended information available in Supplementary Table 2.

General architecture. Phylogenetic analysis shows broad-scale congruence of the major Sarbecovirus clades when comparing the whole genome and S gene topologies (Figure 1b). The exception to this is the clade of bat coronaviruses most closely related to SARS-CoV-2, which are separated by the pangolin and SARS-CoV-2 clade in the S gene tree. This scattering of the bat clade at the S gene level reflects an evolutionary history of frequent viral recombination events between bat coronaviruses. In contrast, the structural similarity of S proteins showed a high level of incongruence when compared to the nucleotide phylogenetic trees. However, the closest relatives to SARS-CoV-2 continue to belong to the pangolin and bat coronavirus clades, consistent with current literature (Holmes, 2024; Liu et al., 2020; Temmam et al., 2022; Ye et al., 2020). Previous structural phylogenetic analysis supports this finding by showing pCoV_GX-P4L as the most conserved S protein to SARS-CoV-2 (Aslam et al., 2023). The lowest nucleotide sequence conservation across the Sarbecoviruses exists in the RBD and NTD, corresponding to loops in the structure which are highly flexible and solvent-exposed (Figure 1c) (Buchanan et al., 2022; Cantoni et al., 2022; Cerutti et al., 2021). These regions display the greatest structural differences; however, their mobile nature may be a confounding factor in the structural phylogenetic tree (Figure 2a). For this reason, statistical confidence in the structure tree (as well as that from Aslam et al 2023) may be improved with the use of molecular dynamic methods outlined by Malik et al 2020 (Malik et al., 2020). The overall structure of S glycoproteins is highly conserved across this subgenus (Figure 2a), with bCoV_PDF-2180 as the most distant member. While root mean square deviations (RMSDs) were calculated for all structures at the whole chain, NTD, RBD, and RBM levels, PDF-2180 has been excluded from visual overlays due to its lack of structural conservation in key areas (i.e. RBM), likely due to its closer relationship to MERS-CoV than SARS-CoV-2(Xiong et al., 2022). The remaining 12 Sarbecovirus structures that have been resolved to date show high conservation with RMSD values of unpruned atoms ranging from 1.4 Å – 2.6 Å (Chain A),

0.46 Å – 1.6 Å (RBD), 0.52 Å – 1.6 Å (RBM), and 1.0 Å – 3.1 Å (NTD) (Supplementary Figure 3).

Generally, the main chain of the RBD is the most conserved while the NTD shows the lowest structural conservation, due to highly variable NTD loops (Buchanan et al., 2022; Cantoni et al., 2022; Klinakis et al., 2021). The RMSD overlays (Figure 2a) and values (Supplementary Figure 1) align with the sequence conservation model (Figure 1b) where the main areas of divergence exist in the NTD and RBM loops, while the majority of structures show consistently high conservation within the main body of the RBD and S1 subunit. In contrast, the protein amino acid sequence similarity of bCoV-BANAL-20-52 shows the highest conservation (98.4%), followed closely by bCoV_RaTG13 (97.4%), while the protein with the highest structural conservation according to S protein structural phylogenetics and RMSD across all domains is pCoV_GX-P4L (Ou et al., 2023; Temmam et al., 2022).

There is complete conservation of disulphide bonds within the NTD and RBD across all solved S proteins (15C-136C, 131C-166C, 291C-301C, 336C-361C, 379C-432C, 391C-525C, 480C-488C following SARS-CoV-2 numbering).

The receptor binding motif (RBM) within SARS-CoV-2 contains five residues that form crucial hydrophilic interactions with hACE2; Y449, Q493, Q498, N501 and Y505, while F486 forms a hydrophobic interaction with hACE2 (SARS-CoV-2 numbering) (Buchanan et al., 2022; Zhang et al., 2021c). Many of the spike proteins reviewed here display variations at one or more of these site, hindering their ability to bind and infect cells expressing hACE2. However, the acquisition of amino acids present in the SARS-CoV-2 spike has been shown to greatly increase the ability of spike proteins from multiple animal host species to bind and infect cells using hACE2. The residue F486 is conserved in pCoV_GD, bCoV_WIV1, bCoV_BANAL-20-52 and 236, while the remaining CoVs possess one of the less bulky Leu or Pro amino acids at this site, conserving the local hydrophobicity (Zhang et al., 2021c; Zhang et al., 2023b). In SARS-CoV-2, Y449 forms hydrogen bonds with hACE2 D38 and Q42 (Buchanan et al., 2022). The equivalent site is conserved in all other S proteins excluding bCoV_RaTG13 and PRD-0038, which possess a Phe (Lee et al., 2023; Zhang et al., 2021c). Introducing a F449Y mutation in bCoV_RaTG13 increases hACE2 binding by ~2-fold (Zhang et al., 2021c). Q493, which potentially interacts with hACE2 K31/E35 is conserved only in pCoV_GD and bCoV_BANAL-20-52, while bCoV_BANAL-20-236, PRD-0038 and cCoV_SZ3 possess a Lys which contacts residues present in bACE2 (Lee et al., 2023). Q498 has been implicated along with Q493 in host-specific ACE2 interactions (Lee et al., 2023; Ou

et al., 2023). While no other S proteins solved possess Q498, pCoV_GD and GX, and bCoV_BANAL-20-52 and 236 all have a His at this site. Research on the host recognition and cell entry of bCoV_BANAL and SARS-CoV-2 shows that introducing the Q498H mutation in SARS-CoV-2 significantly increases pseudovirus entry to cells displaying bACE2, while the H498Q mutation in BANAL-20-52 and 236 significantly decreased pseudovirus cell entry, outlining H498 as an important residue for entering bACE2 displaying cells (Ou et al., 2023). In addition, it was shown that while cCoV-hACE2 binding is low, the introduction of K493N or S498T mutations (SARS-CoV-2 numbering) significantly increases the binding and pseudovirus entry of cCoV_SZ3 to cells expressing hACE2 (Li et al., 2005; Liu et al., 2007). SARS-CoV-2 N501 interacts with a negatively charged area of hACE2 and is conserved in pCoV_GD, bCoV_BANAL-20-52 and 236, and WIV1, while pCoV_GX has a Thr at this site, likely maintaining this interaction (Zhang et al., 2021c). Interestingly, bCoV_RaTG13 possesses an acidic Asp residue that would not favour interactions with the corresponding hydrophobic hACE2 region; this is supported by previous results showing the introduction of the D501N mutation in RaTG13 improves hACE2 binding 9-fold (Zhang et al., 2021c). Prior research also shows that the introduction of V490W mutation (equivalent position) in PRD-0038 allows the bCoV to acquire hACE2 binding (Lee et al., 2023). Residue position Y505 is highly conserved among S proteins with only bCoV_RaTG13 and RsSHCO14 diverging with a His at this site (Lee et al., 2023; Wrobel et al., 2021; Wrobel et al., 2020; Zhang et al., 2021c). The introduction of H505Y in bCoV_RaTG13 has been linked with a ~3-fold increase in both hACE2 and mouse ACE2 binding (Zhang et al., 2021c, Li et al., 2023). Development of the H505Y mutation in spike proteins from various animal coronaviruses, along with the aforementioned RBM mutations, may assist in acquisition of novel host species.

The biliverdin binding pocket (BBP) located in the NTD was shown to contribute to immune escape when occupied with heme metabolite biliverdin (Freeman et al., 2023; Rosa et al., 2021). The ability of SARS-CoV-2 to recruit and bind biliverdin resulted in worse disease outcome in patients (Rosa et al., 2021). Therefore, understanding the prevalence of the heme sequestering across the sarbecovirus subgenus, may better prepare us for the potential pathogenesis of zoonotically emerging human coronaviruses.

While 9 of the 13 structures show some form of a conserved hydrophobic pocket, their width, height, and depth vary. Electrostatic potential maps were assessed for density within NTD hydrophobic pockets to determine whether unmodelled ligands were present (Figure 2b).

Density which may correspond to the heme metabolite biliverdin was detected in 7 of the maps: SARS-CoV-2, cCoV_SZ3 and 007, bCoV-RsSHC014, PRD-0038, WIV1, and pCoV-GX. Both SARS-CoV and bCoV_BANAL-20-236 have a BBP that appears conserved enough to allow ligand binding but does not contain any unassigned density suggestive of a ligand present. While bCoV_PDF-2180 conserved two antiparallel ®-sheets in this region, it shows a complete absence of any hydrophobic pocket. Interestingly, the three remaining proteins possess the same difference in structure that prevents the formation of the BBP, a significant movement in the ⟨-chain, which forms the structure between the two ®-sheets comprising the BBP (SARS-CoV-2: S117-K187). When compared to SARS-CoV-2 the largest distance between the structures at L176 is 9.8 Å (bCoV-BANAL-20-52), 12.5 Å (bCoV-RaTG13), and 12.7 Å (pCoV-GD).

The hydrophobic fatty acid binding pocket (FABP) located next to an RBD antiparallel ®-sheet is occupied by the essential fatty acid, linoleic acid (LA), in SARS-CoV-2 (PDB: 7QUS) with its carboxyl headgroup oriented towards the nearby R408 and Q409 (Figure 2c) (Buchanan et al., 2022). While all S proteins reviewed here maintain the R408 and Q409 equivalent residues in their structure, only five have electrostatic potential maps with evidence significant enough for the authors to model LA within the FABP; SARS-CoV, bCoV_WIV1, cCoV_SZ3 and 007, and pCoV_GX. When overlayed the LA's show very similar areas of occupation and overlap with the electrostatic potential map of 7QUS. Authors have previously commented on the absence of LA in bCoV-RaTG13 despite conservation of key structural residues and concluded further investigation into the mechanism of LA binding was required (Zhang et al., 2021c). Acquiring the ability to bind LA may provide multiple selective advantages by preventing premature 'open' conformation, resulting in a more stable S protein and burying of the RBD and RBM antigenic epitopes (Berger and Schaffitzel, 2020; Qiao and Wang, 2024; Toelzer et al., 2020; Toelzer et al., 2022). However, LA binding also presents an avenue for antiviral treatment of pathogenic coronaviruses and may be effective against emerging zoonotic coronaviruses in the future. SARS-CoV-2 sequesters LA, resulting in the upregulation of cPLA2 activity, an enzyme that assists in coronavirus-induced membrane rearrangements (Toelzer et al., 2022). Treatment with excess LA interferes with virion production by inhibiting cPLA2, which results in the downregulation of membrane remodelling required for viral replication (Toelzer et al., 2022). Prior research on excess LA treatment hindering virion production has also been carried out in MERS-CoV (Yan et al., 2019).

While SARS-CoV-2 is readily found in the 'open' conformation, all the animal coronavirus S proteins investigated preferentially adopt a 'closed' conformation (Hills et al., 2024; Lee et al., 2023; Ou et al., 2023; Qiao and Wang, 2024; Wrobel et al., 2021; Xiong et al., 2022; Zhang et al., 2021c). The furin cleavage site, present in SARS-CoV-2 but in none of the animal coronavirus spike proteins, is thought to be partially responsible for this phenotypic divergence (Berger and Schaffitzel, 2020; Chan and Zhan, 2022; Wrobel et al., 2020). The furin cleavage site in SARS-CoV-2 allows for the early cleavage of the S1-S2 subunits prior to mature virion release. Whilst creating a lower level of protein stability, this early cleavage primes the SARS–CoV-2 spike protein for viral fusion, providing a significant advantage in fusion with the host cell resulting in increased pathogenesis. While being more 'primed' for the 'open' conformation resulting in increased virulence, it's thought that SARS-CoV-2 has offset this lack of stability by forming tighter interactions compared to other S proteins (Berger and Schaffitzel, 2020; Wrobel et al., 2020; Yan et al., 2021).

Interacting surface areas ($Å^2$) of the various S protein chains and trimeric interfaces (central helix) were compared, showing value ranges of ~4700 Å to ~6200 Å and 226 Å to 282 Å respectively (Figure 2d). A tightly packed closed conformation has been previously discussed to increase trimer stability and by extension virulence of SARS-CoV-2 (Xu et al., 2021). A comparison of the S protein buried surface area shows that SARS-CoV-2 has the largest interacting interfaces (6222 Å) while bCoV_PRD-0038 has the smallest (4713 Å). From these observations, we assessed the differences in interactions at the trimeric interface between SARS-CoV-2 and bCoV_PRD-0038 (Figure 2e). While four amino acid side chains are within reasonable contact distance (4 Å) in bCoV_PRD-0038 (Q988, T989, L995, I996), eight are present in SARS-CoV-2 (Q1002, Q1005, T1009, Q1010, I1013, L1012, E1017, R1019). The greater number of interactions at the interface could be a contributing factor to a more tightly bound closed conformation (Walls et al., 2016a; Xu et al., 2021).

The fusion peptide proximal region (FPPR) and 630 loops are partially responsible for the RDB 'up' and 'down' conformation in S proteins (Benton et al., 2020; Cai et al., 2021; Zhang et al., 2021a; Zhang et al., 2021b). These regions are disordered in SARS-CoV-2 indicating a high level of flexibility contributing to more 'open' conformation particles (Berger and Schaffitzel, 2020; Cai et al., 2021; Yang et al., 2021). Meanwhile, bCoV_BANAL-20-52 and 236 have ordered FPPR and 630 loops which have been shown to insert between the SD2 and NTD, stabilising SD2 while also restricting the movement of SD1 (Ou et al., 2023). In

bCoV_RsSHCO14, the introduction of the Y623H mutation causes a ~200-fold increase in pseudovirus entry to cells expressing hACE2 (Qiao and Wang, 2024). Current research suggests that this mutation introduces a charge that may destabilise the SD2 loop and facilitate open conformation, similar to that of mutation D614G in SARS-CoV-2 which is associated with enhanced infectivity (Berger and Schaffitzel, 2020; Dokainish and Sugita, 2023; Zhang et al., 2021a). This is supported by the H623Y mutation in WIV1 causing a reduction in RBD flexibility (Qiao and Wang, 2024).

The RBD conformation. Many single amino acid substitutions have been linked to the preferentially 'closed' conformation of animal CoV S proteins. The NTD L50 residue is present in all S proteins excluding SARS-CoV-2 and causes an NTD rotation which promotes the 'down' conformation (Wrobel et al., 2021). K417 found in SARS-CoV-2, bCoV_RaTG13, BANAL-20-52 and 236 forms a salt bridge with G406 (SARS-CoV-2 numbering) (Amin et al., 2020; Lee et al., 2023). While in pCoV_GX the corresponding residue (R417) retains the Gly salt bridge it also allows stacking interactions with R403 and Y505 of the neighbouring RBD (Lee et al., 2023; Wrobel et al., 2021). Finally, A372 in SARS-CoV-2, which is conserved in every other S protein as T372 (excluding PRD-0038). Previous research shows A372 is favoured in SARS-CoV-2 and the introduction of mutation A372T reduces infectivity by ~20-fold (Kang et al., 2021). When the T372A mutation is introduced to BANAL-20-52 and 236, pseudovirus entry to cells expressing both hACE2 and bACE2 is increased (Ou et al., 2023). T372 allows the conservation of glycosylation at N370 which is implicated in promoting the 'down' conformation and is absent in SARS-CoV-2 (Lee et al., 2023; Ou et al., 2023; Zhang et al., 2022; Zhang et al., 2021c). These factors contribute to a preferred 'down' conformation and are thought to be advantageous in bat coronaviruses due to the extra S protein stability required for the faecal-oral transmission pathway where proteins must avoid dissociation in the low pH of bat stomachs (Ou et al., 2023). Whilst advantageous to bat hosts, the loss of these stabilising factors, following a species jump to an intermediate (or human) host would increase viral tansmissibility due to the preferential 'up' orientation, as seen in SARS-CoV-2.

Glycosylation profile. All S proteins reviewed are heavily glycosylated with a total number of N-linked glycosylations per monomer ranging from 11 – 21. Glycan molecules seen here include N-Acetylglucosamine (NAG), β-D-mannopyranose (BMA), mannose (MAN),

and fucose (FUC). While the most common form of glycosylation is the addition of a single NAG molecule, there is a wide range of highly branched, complex glycan trees present across the proteins. Unsurprisingly, the second most common glycan is the simple NAG-NAG addition, followed by NAG-NAG-BMA (present in 007, SZ3, WIV1, RsSHC014). The less common glycan trees contain a range of NAG, BMA, MAN, and FUC molecules in varying configurations and numbers. While it is interesting to compare the varying configurations of complex glycan trees across S protein models, it's important to note that the trees are solvent-exposed and highly flexible structures. Because of this, the presence of glycan density in electrostatic potential maps is only available at higher resolution and for glycan molecules closest to the S protein main chain. This means that although it appears many S proteins only contain NAG and NAG-NAG composition glycosylations, further modifications may be present but cannot be modelled with confidence. In addition, variation exists in the confidence between authors to model glycan molecules within different types of density. For example, both 7CN4 and 8TC0 contain glycosylation sites (N705 and N119 respectively) which appear to have clear density for modelling but have been left empty. Meanwhile, 8U29 has two complex branched glycan trees (N162-NAG-NAG-BMA-MAN and N230-NAG-NAG-BMA-MAN-MAN-MAN), for which density is present but not as defined as those sites unmodelled in 7CN4 and 8TC0. Therefore, while comparison of the extent and types of glycosylations present in S protein models is interesting it is more useful to compare the general glycosylation coverage and the presence of glycosylations with functional significance.

The glycan distribution across S proteins is fairly conserved throughout key domains with the majority clustering in the NTD and RBD while the rest are spread across the remaining S1 and S2 domains. While the total S protein glycosylation is described as a 'shield' against the host immune response and interact with alternate receptors for increased viral attachment, specific glycosylations (i.e. N343) are involved in both the steric hindrance of antibody binding, contribute to the antigenic epitope (Peng et al., 2021; Chawla et al., 2022; Gong et al., 2021; Walls et al., 2016b; Watanabe et al., 2020; Zhang et al., 2023a). Other glycosites, such as RBD N165, N234, and N370 are implicated in stabilising the open or closed conformation states of S protein by making contact with the neighbouring RBD (Gong et al., 2021; Harbison et al., 2022; Zhang et al., 2022). Glycosylated N165 is present in all S proteins assessed, as is glycosylated N234 with the exception of cCoV_SZ3 and 007 which have not conserved the Asn amino acid. Meanwhile, glycosylated N370 is present in all S proteins except SARS-CoV-

2, due to the previously mentioned T372A adaptation which eliminates the glycosite sequon (Zhang et al., 2022).

Concluding statements and future perspective

In-depth structural comparison of spike glycoproteins closely related to SARS-CoV-2 is essential in understanding the virus-host interactions that drive the evolution of coronaviruses and their tendency to jump species' boundaries and emerge in new hosts. Structural studies of spike glycoproteins suggest that SARS-CoV-2 may have arisen from a recombination event between currently unidentified viruses with RBD sequences and binding properties similar to bCoV_RaTG13 and pCoV_GX due to the similarities SARS-CoV-2 shares with both. Spike proteins compared in this study have previously shown the ability to bind and infect hACE2-presenting cells naturally or with single amino acid mutations. Surveillance studies have also shown evidence of human exposure to multiple animal coronaviruses including bCoV_RaTG13, BANAL52, LYRa11, Rs2018B, RsSHC014, WIV1 and pCOV_GD1, and GX-P5L (Evans et al., 2023). This knowledge, combined with the increase in viral zoonotic emergence events in recent years, further outlines the potential risks animal coronaviruses pose to human health. Spike proteins closely related to SARS-CoV-2 have been structurally investigated from a narrow range of animal hosts, preventing a full understanding of S protein-host receptor interactions and how these evolve. Advancements in this area can be made through the combined use of experimental (cryo-EM/biochemical assays), and computational (AlphaFold, Modeller, RoseTTAFold) techniques to develop a comprehensive understanding of spike proteins from diverse host species providing the fundamental knowledge for the development of broad acting sarbecovirus therapeutics, thus aiding in pandemic preparedness.

Methodology

Sequence-based phylogenetic trees. Maximum likelihood phylogenetic trees were estimated using IQ-Tree v1.6.12 following nucleotide alignment using MAFFT for the full genome and the gene encoding the spike protein (Katoh et al., 2002; Nguyen et al., 2015). Each phylogenetic tree contains 344 sequences obtained from GenBank (http://www.ncbi.nlm.nih.gov) (accession numbers in Supplementary figure 1) (Benson et al., 2013). Trees are mid-point rooted for clarity.

Structure-based phylogenetic tree. Structural similarity dendrogram was produced by the Dali similarity matrix of pairwise Z-scores by average linkage clustering in the 'all against all' Dali structure comparison server (Holm et al., 2023). Branch lengths are modelled ad hoc in the Dali server as the difference in Z-score between structures.

Structural conservation. To perform the sequence conservation structure, sequence alignment was carried out for all the spike glycoprotein genes using Geneious 2024.0.7 global alignment with free end gaps and the Blosum62 cost matrix. Sequence alignment was imported to ChimeraX_Daily (downloaded 13 Sept 2024), where worm representation was applied and coloured by the sequence conservation attribute on the SARS-CoV-2 S protein monomer (PDB: 7QUS) (Pettersen et al., 2021). Structural alignments were carried out using the ChimeraX 1.7.1 matchmaker function with the Needleman-walsh alignment algorithum and best chain pairing. Independent alignments were completed for the monomer (chain A), N-terminal domain (NTD), receptor binding domain (RBD), and receptor binding motif (RBM). All RMSD calculations were reported using unpruned atoms. Models in hydrophobic surface representation were fit to electron microscopy (EM) maps using ChimeraX 1.7.1 map to model fit to compare both the fatty acid binding pocket (FABP) and biliverdin binding pockets (BBP) and potential ligands bound. The Å2 surface area of interacting interfaces was calculated using the PDBePISA server v1.52 (Krissinel and Henrick, 2007).

Structural information. Information relating to the PDB-published structures of spike proteins was collated from the PDB validation reports and original publications where these proteins were first described (Berman et al., 2000). Validation data was obtained by importing the PDB model and map files for each structure into the comprehensive validation job in PHENIX v 1.21-5207 (Liebschner et al., 2019).

Data Availability

GenBank accession codes for genomes included in Figure 1 can be found in Supplementary Table 1 (Benson et al., 2013). The atomic coordinates for structures compared in this review were obtained from the Worldwide Protein Data Bank, accession codes can be found in Supplementary Table 2 (Berman et al., 2000).

Acknowledgements

The authors would like to acknowledge James Hodgkinson-Bean for his technical support.

Conflicting Interests

The authors report no competing interests.

References:

Allen, T., Murray, K.A., Zambrana-Torrelio, C., Morse, S.S., Rondinini, C., Di Marco, M., Breit, N., Olival, K.J., Daszak, P., 2017. Global hotspots and correlates of emerging zoonotic diseases. Nat Commun 8, 1124.

Amin, M., Sorour, M.K., Kasry, A., 2020. Comparing the Binding Interactions in the Receptor Binding Domains of SARS-CoV-2 and SARS-CoV. J Phys Chem Lett 11, 4897-4900.

Aslam, M., Nawaz, M.S., Fournier-Viger, P., Li, W., 2023. Comparative Analysis and Classification of SARS-CoV-2 Spike Protein Structures in PDB. Covid 3, 452-471.

Belouzard, S., Chu, V.C., Whittaker, G.R., 2009. Activation of the SARS coronavirus spike protein via sequential proteolytic cleavage at two distinct sites. Proc Natl Acad Sci U S A 106, 5871-5876.

Benson, D.A., Cavanaugh, M., Clark, K., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J., Sayers, E.W., 2013. GenBank. Nucleic Acids Res 41, D36-42.

Benton, D.J., Wrobel, A.G., Xu, P., Roustan, C., Martin, S.R., Rosenthal, P.B., Skehel, J.J., Gamblin, S.J., 2020. Receptor binding and priming of the spike protein of SARS-CoV-2 for membrane fusion. Nature 588, 327-330.

Berger, I., Schaffitzel, C., 2020. The SARS-CoV-2 spike protein: balancing stability and infectivity. Cell Res 30, 1059-1060.

Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., Bourne, P.E., 2000. The Protein Data Bank. Nucleic Acids Res 28, 235-242.

Bosch, B.J., van der Zee, R., de Haan, C.A., Rottier, P.J., 2003. The coronavirus spike protein is a class I virus fusion protein: structural and functional characterization of the fusion core complex. J Virol 77, 8801-8811.

Buchanan, C.J., Gaunt, B., Harrison, P.J., Yang, Y., Liu, J., Khan, A., Giltrap, A.M., Le Bas, A., Ward, P.N., Gupta, K., Dumoux, M., Tan, T.K., Schimaski, L., Daga, S., Picchiotti, N., Baldassarri, M., Benetti, E., Fallerini, C., Fava, F., Giliberti, A., Koukos, P.I., Davy, M.J., Lakshminarayanan, A., Xue, X., Papadakis, G., Deimel, L.P., Casablancas-Antras, V., Claridge, T.D.W., Bonvin, A., Sattentau, Q.J., Furini, S., Gori, M., Huo, J., Owens, R.J., Schaffitzel, C., Berger, I., Renieri, A., Study, G.-C.M., Naismith, J.H., Baldwin, A.J., Davis, B.G., 2022. Pathogen-sugar interactions revealed by universal saturation transfer analysis. Science 377, eabm3125.

Cai, Y., Zhang, J., Xiao, T., Lavine, C.L., Rawson, S., Peng, H., Zhu, H., Anand, K., Tong, P., Gautam, A., Lu, S., Sterling, S.M., Walsh, R.M., Jr., Rits-Volloch, S., Lu, J., Wesemann, D.R., Yang, W., Seaman, M.S., Chen, B., 2021. Structural basis for enhanced infectivity and immune evasion of SARS-CoV-2 variants. Science 373, 642-648.

Cantoni, D., Murray, M.J., Kalemera, M.D., Dicken, S.J., Stejskal, L., Brown, G., Lytras, S., Coey, J.D., McKenna, J., Bridgett, S., Simpson, D., Fairley, D., Thorne, L.G., Reuschl, A.K., Forrest, C., Ganeshalingham, M., Muir, L., Palor, M., Jarvis, L., Willett, B., Power, U.F., McCoy, L.E., Jolly, C., Towers, G.J., Doores, K.J., Robertson, D.L., Shepherd, A.J., Reeves,

M.B., Bamford, C.G.G., Grove, J., 2022. Evolutionary remodelling of N-terminal domain loops fine-tunes SARS-CoV-2 spike. EMBO Rep 23, e54322.

Cerutti, G., Guo, Y., Zhou, T., Gorman, J., Lee, M., Rapp, M., Reddem, E.R., Yu, J., Bahna, F., Bimela, J., Huang, Y., Katsamba, P.S., Liu, L., Nair, M.S., Rawi, R., Olia, A.S., Wang, P., Zhang, B., Chuang, G.Y., Ho, D.D., Sheng, Z., Kwong, P.D., Shapiro, L., 2021. Potent SARS-CoV-2 neutralizing antibodies directed against spike N-terminal domain target a single supersite. Cell Host Microbe 29, 819-833 e817.

Chan, Y.A., Zhan, S.H., 2022. The Emergence of the Spike Furin Cleavage Site in SARS-CoV-2. Mol Biol Evol 39.

Chawla, H., Fadda, E., Crispin, M., 2022. Principles of SARS-CoV-2 glycosylation. Curr Opin Struct Biol 75, 102402.

Crits-Christoph, A., Levy, J.I., Pekar, J.E., Goldstein, S.A., Singh, R., Hensel, Z., Gangavarapu, K., Rogers, M.B., Moshiri, N., Garry, R.F., Holmes, E.C., Koopmans, M.P.G., Lemey, P., Peacock, T.P., Popescu, S., Rambaut, A., Robertson, D.L., Suchard, M.A., Wertheim, J.O., Rasmussen, A.L., Andersen, K.G., Worobey, M., Debarre, F., 2024. Genetic tracing of market wildlife and viruses at the epicenter of the COVID-19 pandemic. Cell 187, 5468-5482 e5411.

Dokainish, H.M., Sugita, Y., 2023. Structural effects of spike protein D614G mutation in SARS-CoV-2. Biophys J 122, 2910-2920.

Drosten, C., Gunther, S., Preiser, W., van der Werf, S., Brodt, H.R., Becker, S., Rabenau, H., Panning, M., Kolesnikova, L., Fouchier, R.A., Berger, A., Burguiere, A.M., Cinatl, J., Eickmann, M., Escriou, N., Grywna, K., Kramme, S., Manuguerra, J.C., Muller, S., Rickerts, V., Sturmer, M., Vieth, S., Klenk, H.D., Osterhaus, A.D., Schmitz, H., Doerr, H.W., 2003. Identification of a novel coronavirus in patients with severe acute respiratory syndrome. N Engl J Med 348, 1967-1976.

Evans, T.S., Tan, C.W., Aung, O., Phyu, S., Lin, H., Coffey, L.L., Toe, A.T., Aung, P., Aung, T.H., Aung, N.T., Weiss, C.M., Thant, K.Z., Htun, Z.T., Murray, S., Wang, L., Johnson, C.K., Thu, H.M., 2023. Exposure to diverse sarbecoviruses indicates frequent zoonotic spillover in human communities interacting with wildlife. Int J Infect Dis 131, 57-64.

Fan, Y., Zhao, K., Shi, Z.L., Zhou, P., 2019. Bat Coronaviruses in China. Viruses 11.

Freeman, S.L., Oliveira, A.S.F., Gallio, A.E., Rosa, A., Simitakou, M.K., Arthur, C.J., Mulholland, A.J., Cherepanov, P., Raven, E.L., 2023. Heme binding to the SARS-CoV-2 spike glycoprotein. J Biol Chem 299, 105014.

Gong, Y., Qin, S., Dai, L., Tian, Z., 2021. The glycosylation in SARS-CoV-2 and its receptor ACE2. Signal Transduction and Targeted Therapy 6.

Harbison, A.M., Fogarty, C.A., Phung, T.K., Satheesan, A., Schulz, B.L., Fadda, E., 2022. Fine-tuning the spike: role of the nature and topology of the glycan shield in the structure and dynamics of the SARS-CoV-2 S. Chem Sci 13, 386-395.

Hills, F.R., Eruera, A.R., Hodgkinson-Bean, J., Jorge, F., Easingwood, R., Brown, S.H.J., Bouwer, J.C., Li, Y.P., Burga, L.N., Bostina, M., 2024. Variation in structural motifs within SARS-related coronavirus spike proteins. PLoS Pathog 20, e1012158.

Hoffmann, M., Kleine-Weber, H., Schroeder, S., Kruger, N., Herrler, T., Erichsen, S., Schiergens, T.S., Herrler, G., Wu, N.H., Nitsche, A., Muller, M.A., Drosten, C., Pohlmann, S., 2020. SARS-CoV-2 Cell Entry Depends on ACE2 and TMPRSS2 and Is Blocked by a Clinically Proven Protease Inhibitor. Cell 181, 271-280 e278.

Holm, L., Laiho, A., Toronen, P., Salgado, M., 2023. DALI shines a light on remote homologs: One hundred discoveries. Protein Sci 32, e4519.

Holmes, E.C., 2024. The Emergence and Evolution of SARS-CoV-2. Annu Rev Virol 11, 21-42.

Holmes, E.C., Goldstein, S.A., Rasmussen, A.L., Robertson, D.L., Crits-Christoph, A., Wertheim, J.O., Anthony, S.J., Barclay, W.S., Boni, M.F., Doherty, P.C., Farrar, J., Geoghegan, J.L., Jiang, X., Leibowitz, J.L., Neil, S.J.D., Skern, T., Weiss, S.R., Worobey, M., Andersen, K.G., Garry, R.F., Rambaut, A., 2021. The origins of SARS-CoV-2: A critical review. Cell 184, 4848-4856.

Kang, L., He, G., Sharp, A.K., Wang, X., Brown, A.M., Michalak, P., Weger-Lucarelli, J., 2021. A selective sweep in the Spike gene has driven SARS-CoV-2 human adaptation. Cell 184, 4392-4400 e4394.

Katoh, K., Misawa, K., Kuma, K., Miyata, T., 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res 30, 3059-3066.

Klinakis, A., Cournia, Z., Rampias, T., 2021. N-terminal domain mutations of the spike protein are structurally implicated in epitope recognition in emerging SARS-CoV-2 strains. Comput Struct Biotechnol J 19, 5556-5567.

Krissinel, E., Henrick, K., 2007. Inference of macromolecular assemblies from crystalline state. J Mol Biol 372, 774-797.

Kuiken, T., Fouchier, R.A., Schutten, M., Rimmelzwaan, G.F., van Amerongen, G., van Riel, D., Laman, J.D., de Jong, T., van Doornum, G., Lim, W., Ling, A.E., Chan, P.K., Tam, J.S., Zambon, M.C., Gopal, R., Drosten, C., van der Werf, S., Escriou, N., Manuguerra, J.C., Stohr, K., Peiris, J.S., Osterhaus, A.D., 2003. Newly discovered coronavirus as the primary cause of severe acute respiratory syndrome. Lancet 362, 263-270.

Lam, T.T., Jia, N., Zhang, Y.W., Shum, M.H., Jiang, J.F., Zhu, H.C., Tong, Y.G., Shi, Y.X., Ni, X.B., Liao, Y.S., Li, W.J., Jiang, B.G., Wei, W., Yuan, T.T., Zheng, K., Cui, X.M., Li, J., Pei, G.Q., Qiang, X., Cheung, W.Y., Li, L.F., Sun, F.F., Qin, S., Huang, J.C., Leung, G.M., Holmes, E.C., Hu, Y.L., Guan, Y., Cao, W.C., 2020. Identifying SARS-CoV-2-related coronaviruses in Malayan pangolins. Nature 583, 282-285.

Lavie, M., Dubuisson, J., Belouzard, S., 2022. SARS-CoV-2 Spike Furin Cleavage Site and S2' Basic Residues Modulate the Entry Process in a Host Cell-Dependent Manner. J Virol 96, e0047422.

Lee, J., Zepeda, S.K., Park, Y.J., Taylor, A.L., Quispe, J., Stewart, C., Leaf, E.M., Treichel, C., Corti, D., King, N.P., Starr, T.N., Veesler, D., 2023. Broad receptor tropism and immunogenicity of a clade 3 sarbecovirus. Cell Host Microbe 31, 1961-1973 e1911.

Li P, Hu J, Liu Y, Ou X, Mu Z, Lu X, Zan F, Cao M, Tan L, Dong S, Zhou Y, Lu J, Jin Q, Wang J, Wu Z, Zhang Y, Qian Z. 2023. Effect of polymorphism in Rhinolophus affinis ACE2 on entry of SARS-CoV-2 related bat coronaviruses. PLoS Pathog 19:e1011116.

Li, W., Moore, M.J., Vasilieva, N., Sui, J., Wong, S.K., Berne, M.A., Somasundaran, M., Sullivan, J.L., Luzuriaga, K., Greenough, T.C., Choe, H., Farzan, M., 2003. Angiotensin-converting enzyme 2 is a functional receptor for the SARS coronavirus. Nature 426, 450-454.

Li, W., Zhang, C., Sui, J., Kuhn, J.H., Moore, M.J., Luo, S., Wong, S.K., Huang, I.C., Xu, K., Vasilieva, N., Murakami, A., He, Y., Marasco, W.A., Guan, Y., Choe, H., Farzan, M., 2005. Receptor and viral determinants of SARS-coronavirus adaptation to human ACE2. EMBO J 24, 1634-1643.

Liebschner, D., Afonine, P.V., Baker, M.L., Bunkoczi, G., Chen, V.B., Croll, T.I., Hintze, B., Hung, L.W., Jain, S., McCoy, A.J., Moriarty, N.W., Oeffner, R.D., Poon, B.K., Prisant, M.G., Read, R.J., Richardson, J.S., Richardson, D.C., Sammito, M.D., Sobolev, O.V., Stockwell, D.H., Terwilliger, T.C., Urzhumtsev, A.G., Videau, L.L., Williams, C.J., Adams, P.D., 2019.

Macromolecular structure determination using X-rays, neutrons and electrons: recent developments in Phenix. Acta Crystallogr D Struct Biol 75, 861-877.

Liu, L., Fang, Q., Deng, F., Wang, H., Yi, C.E., Ba, L., Yu, W., Lin, R.D., Li, T., Hu, Z., Ho, D.D., Zhang, L., Chen, Z., 2007. Natural mutations in the receptor binding domain of spike glycoprotein determine the reactivity of cross-neutralization between palm civet coronavirus and severe acute respiratory syndrome coronavirus. J Virol 81, 4694-4700.

Liu, Y., Hu, G., Wang, Y., Ren, W., Zhao, X., Ji, F., Zhu, Y., Feng, F., Gong, M., Ju, X., Zhu, Y., Cai, X., Lan, J., Guo, J., Xie, M., Dong, L., Zhu, Z., Na, J., Wu, J., Lan, X., Xie, Y., Wang, X., Yuan, Z., Zhang, R., Ding, Q., 2021. Functional and genetic analysis of viral receptor ACE2 orthologs reveals a broad potential host range of SARS-CoV-2. Proc Natl Acad Sci U S A 118.

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.

Malik, A.J., Poole, A.M., Allison, J.R., 2020. Structural Phylogenetics with Confidence. Mol Biol Evol 37, 2711-2726.

Menachery, V.D., Yount, B.L., Jr., Debbink, K., Agnihothram, S., Gralinski, L.E., Plante, J.A., Graham, R.L., Scobey, T., Ge, X.Y., Donaldson, E.F., Randell, S.H., Lanzavecchia, A., Marasco, W.A., Shi, Z.L., Baric, R.S., 2015. A SARS-like cluster of circulating bat coronaviruses shows potential for human emergence. Nat Med 21, 1508-1513.

Mittal, A., Khattri, A., Verma, V., 2022. Structural and antigenic variations in the spike protein of emerging SARS-CoV-2 variants. PLoS Pathog 18, e1010260.

Nguyen, L.T., Schmidt, H.A., von Haeseler, A., Minh, B.Q., 2015. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol Biol Evol 32, 268-274.

Ou, X., Liu, Y., Lei, X., Li, P., Mi, D., Ren, L., Guo, L., Guo, R., Chen, T., Hu, J., Xiang, Z., Mu, Z., Chen, X., Chen, J., Hu, K., Jin, Q., Wang, J., Qian, Z., 2020. Characterization of spike glycoprotein of SARS-CoV-2 on virus entry and its immune cross-reactivity with SARS-CoV. Nat Commun 11, 1620.

Ou, X., Xu, G., Li, P., Liu, Y., Zan, F., Liu, P., Hu, J., Lu, X., Dong, S., Zhou, Y., Mu, Z., Wu, Z., Wang, J., Jin, Q., Liu, P., Lu, J., Wang, X., Qian, Z., 2023. Host susceptibility and structural and immunological insight of S proteins of two SARS-CoV-2 closely related bat coronaviruses. Cell Discov 9, 78.

Peng, R., Wu, L.-A., Wang, Q., Qi, J., & Gao, G. F. (2021). Cell entry by SARS-CoV-2. *Trends in biochemical sciences*, *46*(10), 848-860.

Pettersen, E.F., Goddard, T.D., Huang, C.C., Meng, E.C., Couch, G.S., Croll, T.I., Morris, J.H., Ferrin, T.E., 2021. UCSF ChimeraX: Structure visualization for researchers, educators, and developers. Protein Sci 30, 70-82.

Qiao, S., Wang, X., 2024. Structural determinants of spike infectivity in bat SARS-like coronaviruses RsSHC014 and WIV1. J Virol 98, e0034224.

Rahman, M.T., Sobur, M.A., Islam, M.S., Ievy, S., Hossain, M.J., El Zowalaty, M.E., Rahman, A.T., Ashour, H.M., 2020. Zoonotic Diseases: Etiology, Impact, and Control. Microorganisms 8.

Rosa, A., Pye, V.E., Graham, C., Muir, L., Seow, J., Ng, K.W., Cook, N.J., Rees-Spear, C., Parker, E., Dos Santos, M.S., Rosadas, C., Susana, A., Rhys, H., Nans, A., Masino, L., Roustan, C., Christodoulou, E., Ulferts, R., Wrobel, A.G., Short, C.E., Fertleman, M., Sanders, R.W., Heaney, J., Spyer, M., Kjaer, S., Riddell, A., Malim, M.H., Beale, R., MacRae, J.I., Taylor, G.P., Nastouli, E., van Gils, M.J., Rosenthal, P.B., Pizzato, M., McClure, M.O., Tedder, R.S.,

Kassiotis, G., McCoy, L.E., Doores, K.J., Cherepanov, P., 2021. SARS-CoV-2 can recruit a heme metabolite to evade antibody immunity. Sci Adv 7.

Shang, J., Ye, G., Shi, K., Wan, Y., Luo, C., Aihara, H., Geng, Q., Auerbach, A., Li, F., 2020. Structural basis of receptor recognition by SARS-CoV-2. Nature 581, 221-224.

Stadler, K., Masignani, V., Eickmann, M., Becker, S., Abrignani, S., Klenk, H.D., Rappuoli, R., 2003. SARS--beginning to understand a new virus. Nat Rev Microbiol 1, 209-218.

Temmam, S., Vongphayloth, K., Baquero, E., Munier, S., Bonomi, M., Regnault, B., Douangboubpha, B., Karami, Y., Chretien, D., Sanamxay, D., Xayaphet, V., Paphaphanh, P., Lacoste, V., Somlor, S., Lakeomany, K., Phommavanh, N., Perot, P., Dehan, O., Amara, F., Donati, F., Bigot, T., Nilges, M., Rey, F.A., van der Werf, S., Brey, P.T., Eloit, M., 2022. Bat coronaviruses related to SARS-CoV-2 and infectious for human cells. Nature 604, 330-336.

Toelzer, C., Gupta, K., Yadav, S.K.N., Borucu, U., Davidson, A.D., Kavanagh Williamson, M., Shoemark, D.K., Garzoni, F., Staufer, O., Milligan, R., Capin, J., Mulholland, A.J., Spatz, J., Fitzgerald, D., Berger, I., Schaffitzel, C., 2020. Free fatty acid binding pocket in the locked structure of SARS-CoV-2 spike protein. Science 370, 725-730.

Toelzer, C., Gupta, K., Yadav, S.K.N., Hodgson, L., Williamson, M.K., Buzas, D., Borucu, U., Powers, K., Stenner, R., Vasileiou, K., Garzoni, F., Fitzgerald, D., Payre, C., Gautam, G., Lambeau, G., Davidson, A.D., Verkade, P., Frank, M., Berger, I., Schaffitzel, C., 2022. The free fatty acid-binding pocket is a conserved hallmark in pathogenic beta-coronavirus spike proteins from SARS-CoV to Omicron. Sci Adv 8, eadc9179.

Walls, A.C., Park, Y.J., Tortorici, M.A., Wall, A., McGuire, A.T., Veesler, D., 2020. Structure, Function, and Antigenicity of the SARS-CoV-2 Spike Glycoprotein. Cell 181, 281-292 e286.

Walls, A.C., Tortorici, M.A., Bosch, B.J., Frenz, B., Rottier, P.J.M., DiMaio, F., Rey, F.A., Veesler, D., 2016a. Cryo-electron microscopy structure of a coronavirus spike glycoprotein trimer. Nature 531, 114-117.

Walls, A.C., Tortorici, M.A., Frenz, B., Snijder, J., Li, W., Rey, F.A., DiMaio, F., Bosch, B.J., Veesler, D., 2016b. Glycan shield and epitope masking of a coronavirus spike protein observed by cryo-electron microscopy. Nat Struct Mol Biol 23, 899-905.

Watanabe, Y., Allen, J.D., Wrapp, D., McLellan, J.S., Crispin, M., 2020. Site-specific glycan analysis of the SARS-CoV-2 spike. Science 369, 330-333.

Wrapp, D., Wang, N., Corbett, K.S., Goldsmith, J.A., Hsieh, C.L., Abiona, O., Graham, B.S., McLellan, J.S., 2020. Cryo-EM structure of the 2019-nCoV spike in the prefusion conformation. Science 367, 1260-1263.

Wrobel, A.G., Benton, D.J., Xu, P., Calder, L.J., Borg, A., Roustan, C., Martin, S.R., Rosenthal, P.B., Skehel, J.J., Gamblin, S.J., 2021. Structure and binding properties of Pangolin-CoV spike glycoprotein inform the evolution of SARS-CoV-2. Nature Communications 12.

Wrobel, A.G., Benton, D.J., Xu, P., Roustan, C., Martin, S.R., Rosenthal, P.B., Skehel, J.J., Gamblin, S.J., 2020. SARS-CoV-2 and bat RaTG13 spike glycoprotein structures inform on virus evolution and furin-cleavage effects. Nat Struct Mol Biol 27, 763-767.

Xiong, Q., Cao, L., Ma, C., Tortorici, M.A., Liu, C., Si, J., Liu, P., Gu, M., Walls, A.C., Wang, C., Shi, L., Tong, F., Huang, M., Li, J., Zhao, C., Shen, C., Chen, Y., Zhao, H., Lan, K., Corti, D., Veesler, D., Wang, X., Yan, H., 2022. Close relatives of MERS-CoV in bats use ACE2 as their functional receptors. Nature 612, 748-757.

Xu, C., Wang, Y., Liu, C., Zhang, C., Han, W., Hong, X., Wang, Y., Hong, Q., Wang, S., Zhao, Q., Wang, Y., Yang, Y., Chen, K., Zheng, W., Kong, L., Wang, F., Zuo, Q., Huang, Z., Cong, Y., 2021. Conformational dynamics of SARS-CoV-2 trimeric spike glycoprotein in complex with receptor ACE2 revealed by cryo-EM. Sci Adv 7.

Yan, B., Chu, H., Yang, D., Sze, K.H., Lai, P.M., Yuan, S., Shuai, H., Wang, Y., Kao, R.Y., Chan, J.F., Yuen, K.Y., 2019. Characterization of the Lipidomic Profile of Human Coronavirus-Infected Cells: Implications for Lipid Metabolism Remodeling upon Coronavirus ßReplication. Viruses 11.

Yan, R., Zhang, Y., Li, Y., Ye, F., Guo, Y., Xia, L., Zhong, X., Chi, X., Zhou, Q., 2021. Structural basis for the different states of the spike protein of SARS-CoV-2 in complex with ACE2. Cell Res 31, 717-719.

Yang, Z., Han, Y., Ding, S., Shi, W., Zhou, T., Finzi, A., Kwong, P.D., Mothes, W., Lu, M., 2021. SARS-CoV-2 Variants Increase Kinetic Stability of Open Spike Conformations as an Evolutionary Strategy. mBio 13, e0322721.

Ye, Z.W., Yuan, S., Yuen, K.S., Fung, S.Y., Chan, C.P., Jin, D.Y., 2020. Zoonotic origins of human coronaviruses. Int J Biol Sci 16, 1686-1697.

Zaki, A.M., van Boheemen, S., Bestebroer, T.M., Osterhaus, A.D., Fouchier, R.A., 2012. Isolation of a novel coronavirus from a man with pneumonia in Saudi Arabia. N Engl J Med 367, 1814-1820.

Zhang, F., Schmidt, F., Muecksch, F., Wang, Z., Gazumyan, A., Nussenzweig, M.C., Gaebler, C., Caskey, M., Hatziioannou, T., Bieniasz, P.D., 2023a. SARS-CoV-2 spike glycosylation affects function and neutralization sensitivity. bioRxiv.

Zhang, J., Cai, Y., Xiao, T., Lu, J., Peng, H., Sterling, S.M., Walsh, R.M., Jr., Rits-Volloch, S., Zhu, H., Woosley, A.N., Yang, W., Sliz, P., Chen, B., 2021a. Structural impact on SARS-CoV-2 spike protein by D614G substitution. Science 372, 525-530.

Zhang, J., Xiao, T., Cai, Y., Chen, B., 2021b. Structure of SARS-CoV-2 spike protein. Curr Opin Virol 50, 173-182.

Zhang, S., Liang, Q., He, X., Zhao, C., Ren, W., Yang, Z., Wang, Z., Ding, Q., Deng, H., Wang, T., Zhang, L., Wang, X., 2022. Loss of Spike N370 glycosylation as an important evolutionary event for the enhanced infectivity of SARS-CoV-2. Cell Res 32, 315-318.

Zhang, S., Qiao, S., Yu, J., Zeng, J., Shan, S., Tian, L., Lan, J., Zhang, L., Wang, X., 2021c. Bat and pangolin coronavirus spike glycoprotein structures provide insights into SARS-CoV-2 evolution. Nat Commun 12, 1607.

Zhang, W., Shi, K., Geng, Q., Herbst, M., Wang, M., Huang, L., Bu, F., Liu, B., Aihara, H., Li, F., 2023b. Structural evolution of SARS-CoV-2 omicron in human receptor recognition. J Virol 97, e0082223.

Zhao, J., Cui, W., Tian, B.P., 2020. The Potential Intermediate Hosts for SARS-CoV-2. Front Microbiol 11, 580137.

Zhou, P., Yang, X.L., Wang, X.G., Hu, B., Zhang, L., Zhang, W., Si, H.R., Zhu, Y., Li, B., Huang, C.L., Chen, H.D., Chen, J., Luo, Y., Guo, H., Jiang, R.D., Liu, M.Q., Chen, Y., Shen, X.R., Wang, X., Zheng, X.S., Zhao, K., Chen, Q.J., Deng, F., Liu, L.L., Yan, B., Zhan, F.X., Wang, Y.Y., Xiao, G.F., Shi, Z.L., 2020. A pneumonia outbreak associated with a new coronavirus of probable bat origin. Nature 579, 270-273.

Figure Legends

Figure 1. Phylogenetic and structural tree comparison of spike glycoproteins.
(a) (left) SARS-CoV-2 (PDB: 7QUS) spike glycoprotein trimer (pink) in ribbon representation with key domains coloured in purple (N-terminal domain), mint (receptor binding domain), green (fusion protein), yellow (heptad repeat 1). (right) Zoomed in depiction of individual domains annotated above to indicate belonging with either subunit S1 or S2 (b) maximum likelihood phylogenetic trees estimated using IQ-Tree (Nguyen et al., 2015) following nucleotide alignment using MAFFT (Katoh et al., 2002) for the full genome and the gene encoding the spike protein. Each phylogenetic tree contains 344 sequences obtained from GenBank (accession numbers in Supplementary Figure 1). Trees are mid-point rooted for clarity. Structural similarity dendrogram of 13 similar spike protein structures (top to bottom PDB: 7CN8, 7QUS, 7CN4, 8HXJ, 8TC0, 8TC1, 7ZH1, 8TC5, 8U29, 7BBH, 8IW3, 8WLU, 7U6R). Structural phylogeny generated from the Dali similarity matrix of pairwise Z-scores by average linkage clustering. Branch lengths were modelled ad hoc as the difference in Z-scores between pairwise structures. (c) SARS-CoV-2 structure coloured by conservation of sequence within solved coronavirus spike proteins. Geneious multiple sequence alignment was performed with free end gaps and 65% similarity cost matrix. Colouring by conservation performed with ChimeraX, an increase in pink colour shows an increase in conservation.

Figure 2. Structural comparison of functionally relevant domains of sarbecovirus spike proteins.
(a) Structural alignment of spike protein monomer, followed by structural alignment of receptor binding domains (RBD) and receptor binding motifs (RBM) from SARS-CoV-2 (magenta)(PDB:7QUS), SARS-CoV (brown)(PDB:7ZH1), bCoV_RaTG13 (dark blue)(PDB:7CN4), bCoV_WIV1 (yellow)(PDB:8TC0), bCoV_BANAL-20-52 (light purple)(PDB:8HXJ), bCoV_BANAL-20-236 (powder blue)(PDB:8I3W), bCoV_RsSHC014 (dark purple)(PDB:8WLU), bCoV_PRD-0038 (grey)(PDB:8U29), bCoV_PDF-2180 (green)(PDB:7U6R), cCoV_SZ3 (light orange)(PDB:8TC5), cCoV_007 (dark orange)(PDB:8TC1), pCoV_GD (light pink)(PDB:7BBH), pCoV_GX (light blue)(PDB:7CN8). (b) Hydrophobicity surface representation of N-terminal domain hydrophobic biliverdin binding pocket (BBP). Unassigned coulombic potential density within the BBP of SARS-CoV-2, cCoV_SZ3, cCoV_007, bCoV_RsSHC014, bCoV_PDF-2180, bCoV_WIV1, pCoV_GX. (c) Fatty acid binding pocket of SARS-CoV-2 containing linoleic acid (EIC)(light pink) shown in hydrophobicity surface representation with coulombic density shown in mesh representation (blue). Structural alignment of linoleic acids present in solved sarbecovirus spike proteins, SARS-CoV (brown), bCoV_WIV1 (yellow), cCoV_SZ3 (light orange), cCoV_007 (dark orange), pCoV_GX (grey). All residues from SARS-CoV-2 interacting (<4 Å distances) with linoleic acid within the fatty acid binding pocket shown in atom representation. (d) Cartoon representation of SARS-CoV-2 (pink) and bCoV_PRD-0038 (grey) spike trimers followed by the central helix trimeric interface, with interacting residues (<4 Å distances) shown in atomic representation. (e) Interaction interface ($Å^2$) of each spike protein monomer and central helix calculated by PISA.

673

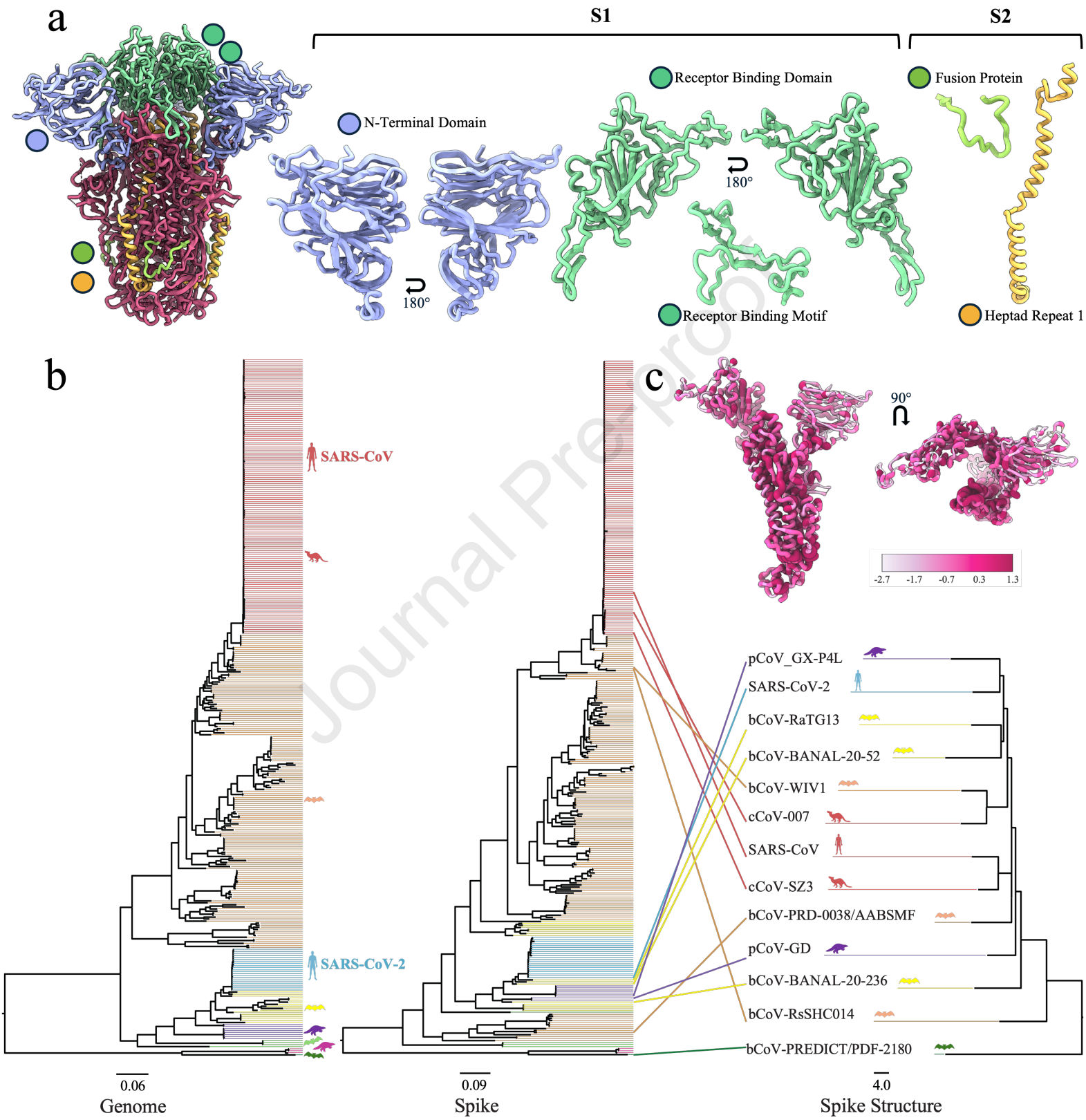
674 **Figure 3. Glycosylation patterns in sarbecovirus spike proteins solved through Cryo-EM.**
675 (a) Schematic of the SARS-CoV-2 spike glycoprotein gene with subunits S1 and S2 and their
676 key domains. Structurally solved glycosylation sites are indicated with a black line and
677 corresponding residue number. N-terminal domain (NTD), receptor binding domain (RBD),
678 fusion peptide (FP), heptad repeat 1 (HR1), heptad repeat 2 (HR2), transmembrane domain
679 (TM), (CP). (b) Example of each type of glycosylation present in spike protein structures, from
680 left to right; NAG (SARS-CoV)(PDB:7ZH1), NAG-NAG (SARS-CoV-2)(PDB:7QUS),
681 NAG-FUC-NAG (cCoV_SZ3)(PDB:8TC5), NAG-NAG-BMA (cCoV_007)(PDB:8TC1),
682 NAG-FUC-NAG-BMA (bCoV_WIV1)(PDB:8TC0), NAG-FUC-NAG-BMA-FUC
683 (cCoV_SZ3)(PDB:8TC5), NAG-NAG-NAG-FUC-BMA-BMA (bCoV_PRD-
684 0038)(PDB:8U29). (c) Cartoon representation of solved sarbecovirus spike protein monomers
685 with modelled glycosylations shown in atomic sphere representations. SARS-CoV-2 (magenta),
686 SARS-CoV (brown), bCoV_RaTG13 (dark blue)(PDB:7CN4), bCoV_WIV1 (yellow),
687 bCoV_BANAL-20-52 (light purple)(PDB(8HXJ), bCoV_BANAL-20-236 (powder
688 blue)(PDB:8I3W), bCoV_RsSHC014 (dark purple*)(8WLU), bCoV_PRD-0038*
689 *(grey)(PDB:8U29), bCoV_PDF-2180 (green)(PDB:7U6R), cCoV_SZ3 (light orange),*
690 *cCoV_007 (dark orange), pCoV_GD (light pink)(PDB:7BBH), pCoV_GX (light blue)(7CN8).*

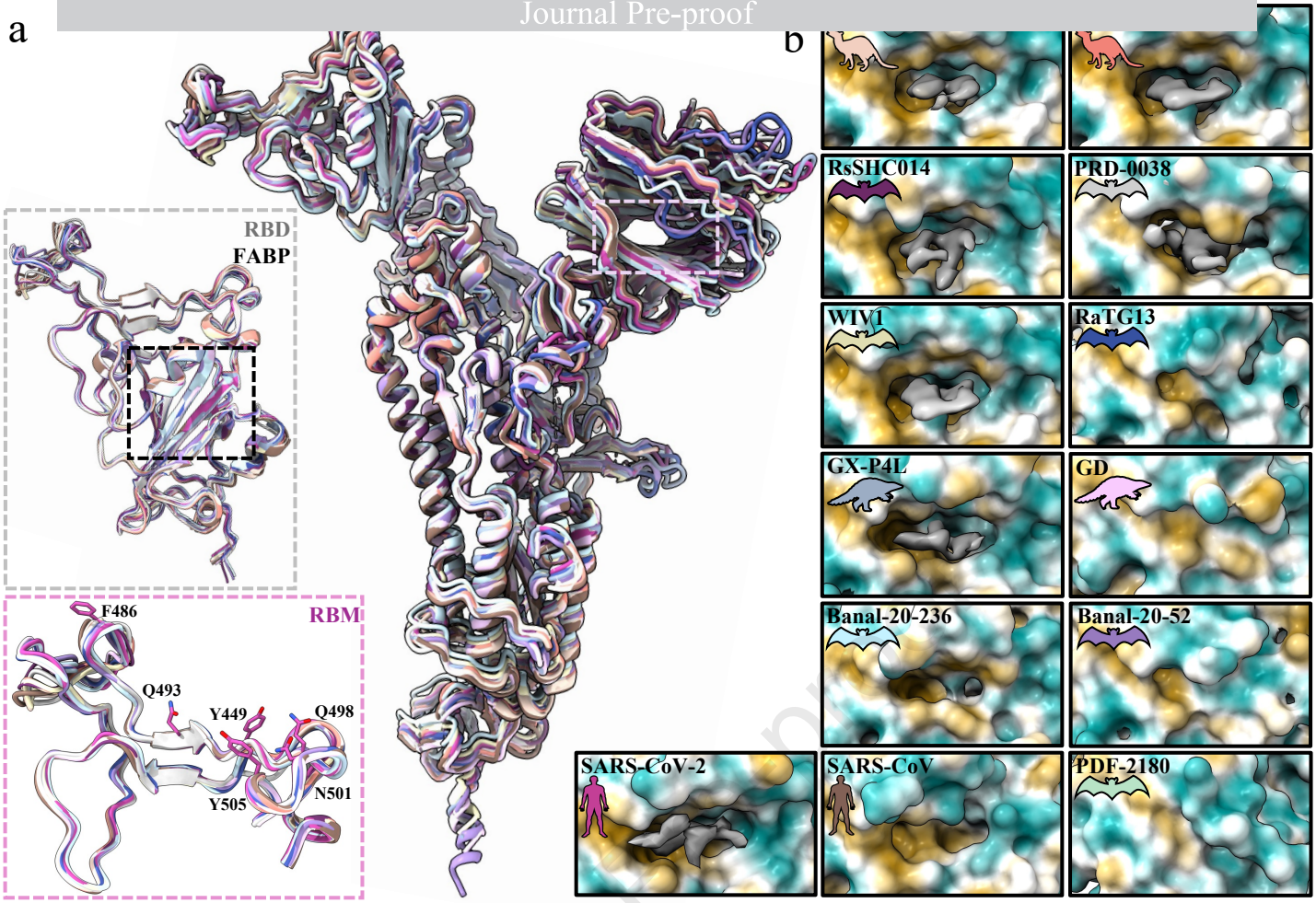
691

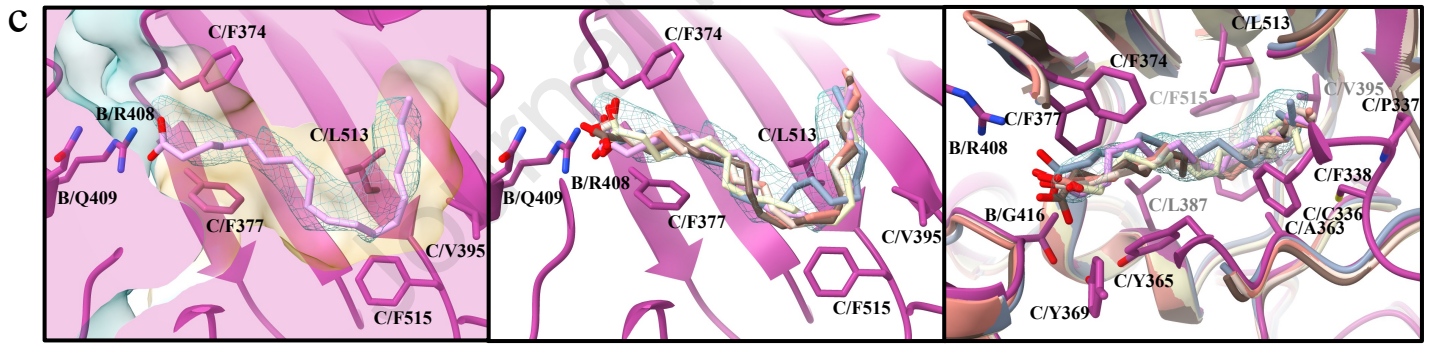
692

Virus name	Virus species	Host species	Methodology	Resolution	Symmetry imposed	Clashscore residues	Clashscore score	Ramachandran plot: Favoured/Allowed/ Disallowed	Refs
SARS-CoV-2 (PDB: 7QUS)	SARS-related	Human	Cryo-EM	2.40 Å	C3	3303	1.68	97.57 2.43 0	Buchanan et al., 2022
SARS-CoV (PDB: 7ZH1)	SARS-related	Human	Cryo-EM	2.48 Å	C3	3060	1.89	96.57 3.43 0	Toelzer et al., 2022
Banal-20-236 (PDB: 8I3W)	Unclassified	Rhinolophus bat	Cryo-EM	2.85 Å	C3	3345	2.11	92.00 8.00 0.00	Ou et al., 2023
Banal-20-52 (8HXJ)	Unclassified	Rhinolophus bat	Cryo-EM	3.52 Å	C3	3399	2.13	90.60 9.40 0.00	Ou et al., 2023
RaTG13 (7CN4)	SARS-related	Rhinolophus bat	Cryo-EM	2.93 Å	C3	3360	1.89	92.70 6.95 0.36	Zhang et al., 2021
WIV1 (8TC0)	SARS-related	Rhinolophus bat	Cryo-EM	1.88 Å	C3	3273	1.4	96.99 3.01 0.00	Hills et al., 2024
RsSHCO14 (PDB: 8WLU)	SARS-related	Rhinolophus bat	Cryo-EM	2.77 Å	C3	3318	1.67	97.49 2.51 0.00	Qiao et al., 2024
PRD-0038 (PDB: 8U29)	SARS-related	Rhinolophus bat	Cryo-EM	2.80 Å	C3	3246	1.26	95.45 4.37 0.19	Lee et al., 2023
PDF-2180 (PDB: 7U6R)	Unclassified	Pipistrellus bat	Cryo-EM	2.50 Å	C3	3291	1.2	98.23 1.58 0.19	Xiong et al., 2022
SZ3 (PDB: 8TC5)	SARS-related	Masked civet	Cryo-EM	2.11 Å	C3	3261	1.22	96.95 3.02 0.03	Hills et al., 2024
007 (PDB: 8TC1)	SARS-related	Masked civet	Cryo-EM	1.92 Å	C3	3264	1.15	97.44 2.56 0.00	Hills et al., 2024
GD (PDB: 7BBH)	Unclassified	Pangolin	Cryo-EM	2.90 Å	C3	3189	1.38	96.13 3.87 0.00	Wrobel et al., 2021
GX-P4L (PDB: 7CN8)	Unclassified	Pangolin	Cryo-EM	2.50 Å	C3	3375	2.07	96.05 3.53 0.42	Zhang et al., 2021

Table 1. Technical information relating to spike glycoprotein structures analysed

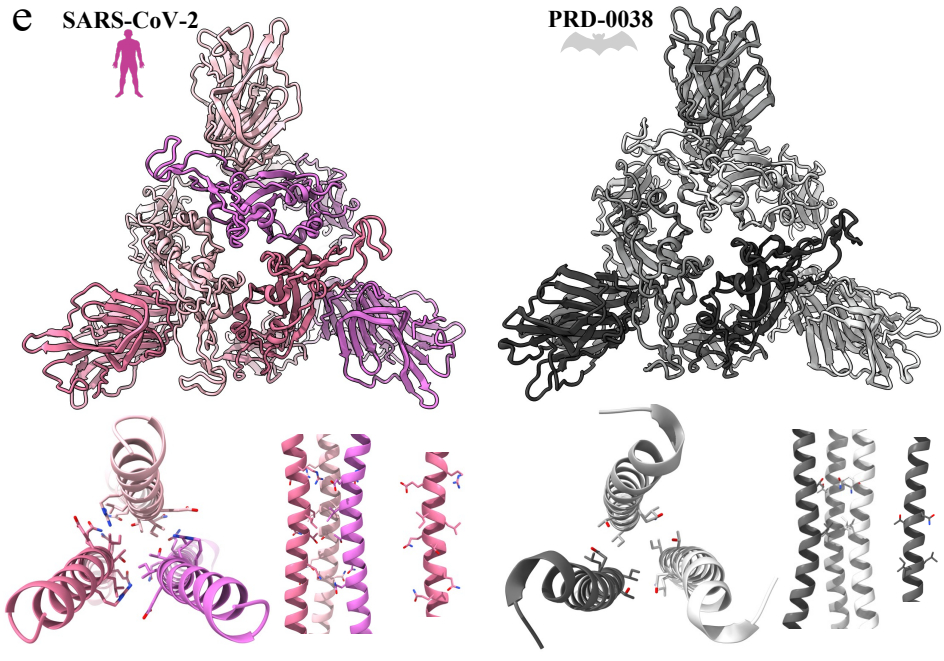


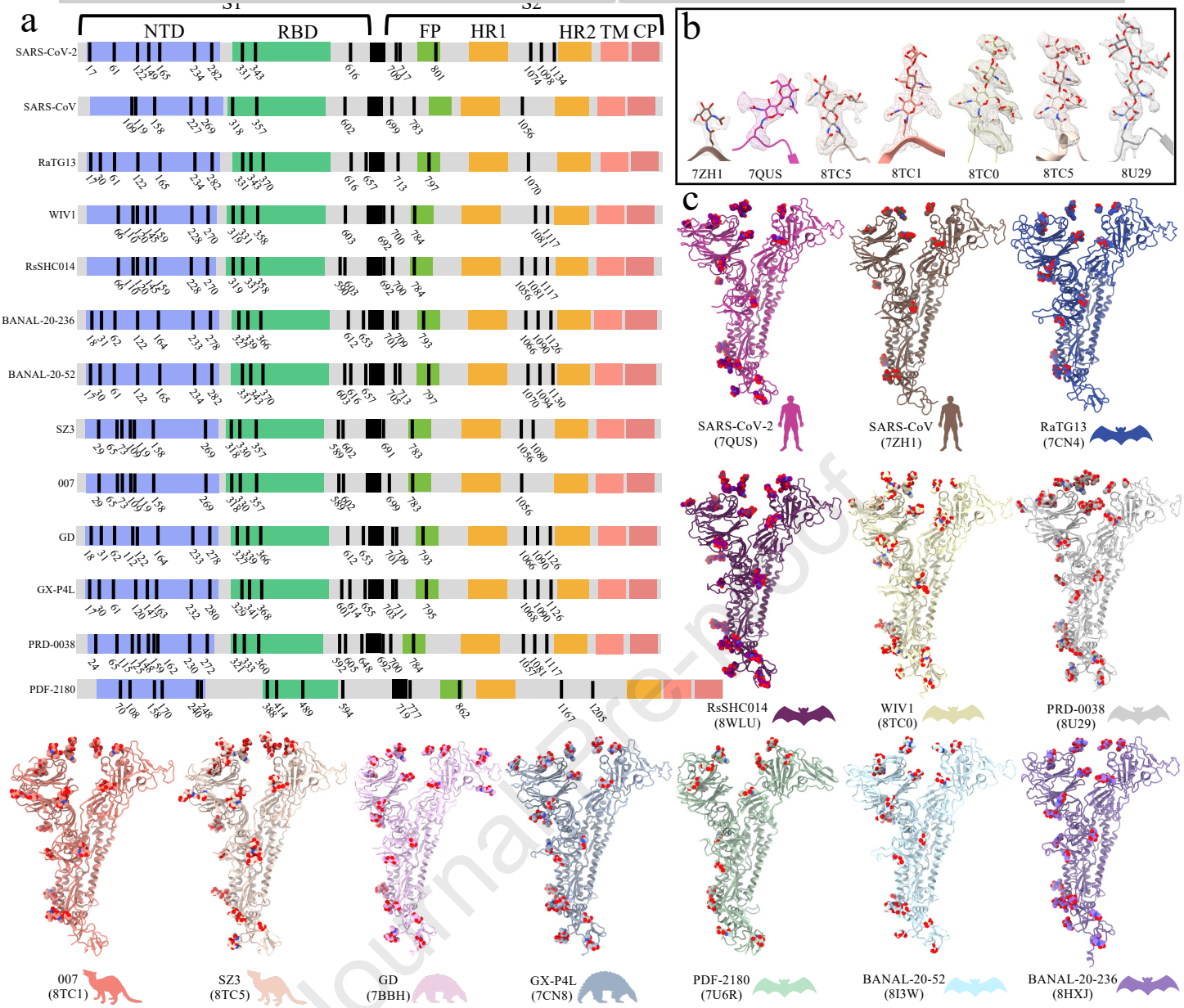




d

Interaction interface (Å²)		
S protein	Monomer	Central Helix
SARS-CoV-2	6222.2	257.6
SARS-CoV	5556.8	226.0
BANAL-20-52	6138.8	262.9
BANAL-20-236	6053.7	268.2
PRD-0038	4713.0	249.2
WIV1	5867.3	246.7
007	5931.3	270.0
SZ3	5814.4	249.8
RsSHC014	5617.8	282.0
GX-P4L	6146.2	243.5
GD	5763.6	257.5
PDF-2180	5039.6	271.0
RaTG13	5827.3	288.1





Review highlights

- Overall structural comparison of spike glycoproteins similar to SARS-CoV-2 reveals high levels of structural conservation and outlines potential host-jumping pathways.
- Many solved spike proteins can perform in vivo human ACE2 binding and pseudovirus entry, which is increased up to ~200-fold with single RBD mutations.
- Sequence and structure variation exists in S proteins similar to SARS-CoV-2 to preferentially adopt the RBD 'down' conformation, favourable for bat coronavirus transmission routes.

Declaration of interests

☒ The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

☐ The author is an Editorial Board Member/Editor-in-Chief/Associate Editor/Guest Editor for *[Journal name]* and was not involved in the editorial review or the decision to publish this article.

☐ The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: