Chemical Science

EDGE ARTICLE

Check for updates

Cite this: *Chem. Sci.*, 2021, **12**, 13492

All publication charges for this article have been paid for by the Royal Society of Chemistry

Modelling the active SARS-CoV-2 helicase complex as a basis for structure-based inhibitor design†

Dénes Berta, ‡[ab] Magd Badaoui, ‡[ab] Sam Alexander Martino, [ab] Pedro J. Buigues, [ab] Andrei V. Pisliakov, *[c] Nadia Elghobashi-Meinhardt, *[d] Geoff Wells, *[e] Sarah A. Harris, *[f] Elisa Frezza *[g] and Edina Rosta *[ab]

Received 20th May 2021
Accepted 1st September 2021

DOI: 10.1039/d1sc02775a

rsc.li/chemical-science

The RNA helicase (non-structural protein 13, NSP13) of SARS-CoV-2 is essential for viral replication, and it is highly conserved among the *coronaviridae* family, thus a prominent drug target to treat COVID-19. We present here structural models and dynamics of the helicase in complex with its native substrates based on thorough analysis of homologous sequences and existing experimental structures. We performed and analysed microseconds of molecular dynamics (MD) simulations, and our model provides valuable insights to the binding of the ATP and ssRNA at the atomic level. We identify the principal motions characterising the enzyme and highlight the effect of the natural substrates on this dynamics. Furthermore, allosteric binding sites are suggested by our pocket analysis. Our obtained structural and dynamical insights are important for subsequent studies of the catalytic function and for the development of specific inhibitors at our characterised binding pockets for this promising COVID-19 drug target.

Introduction

Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) is the causative agent of COVID-19 disease and is responsible for the largest modern pandemic. The virus is closely related to SARS-CoV and MERS-CoV that caused smaller outbreaks of disease earlier this century.[1] Currently, only a few approved drugs have been repurposed for the disease.[2] The approved treatments can be categorized into three groups: aiding respiration in severe cases, repurposed antiviral drugs and monoclonal antibodies. Remdesivir has the longest history of use against COVID-19 infection and shown to bind to the RNA dependent RNA polymerase (RdRp),[3,4] although recent trial data found no evidence for improvement in patient conditions.[5] Further efforts include trials of antiviral combination therapies,[6] or use of the anti-leprosy drug clofazimine, the latter has been found to inhibit helicase activity.[7] However, there is a need for development of specific compounds that can be used to inhibit viral replication for the treatment of COVID-19, prophylaxis of vulnerable individuals and to add to the repertoire of treatment for future coronavirus outbreaks.

Here we focus on determining the structure of catalytically active complexes of the SARS-CoV-2 RNA helicase, also known as non-structural protein 13 (NSP13) (Fig. 1). This protein is part of the Orf1ab polyprotein, that is spliced to produce the enzymes required for viral replication. The RNA helicase performs two essential functions for the viral replication making it an ideal drug target. It is thought to perform the first step in the 5′-capping of the viral RNA by its triphosphatase function hydrolysing the 5′-triphosphate group to form diphosphate-RNA.[8,9] Furthermore, its main helicase function enables RNA translocation and unwinding in an ATP-dependent mechanism during viral replication.

Accordingly, numerous studies have already demonstrated that it is possible to develop potent inhibitors of viral helicases as antiviral agents.[10] The 2003 SARS epidemic inspired a wave of drug development, often in conjunction with other positive RNA viral targets such as the hepatitis C virus (HCV).[11–14] Consequently, HCV helicase inhibitory aryl diketoacids (ADKs) were found useful against SARS-CoV,[11,12,15] in addition to porphyrin metal complexes,[16] and natural[14] or synthetic products.[13,17–19] Typically, these compounds inhibit both the unwinding and the

[a] Department of Physics and Astronomy, University College London, London, WC1E 6BT, UK. E-mail: e.rosta@ucl.ac.uk

[b] Department of Chemistry, King's College London, London, SE1 1DB, UK

[c] Computational Biology, School of Science and Engineering & School of Life Sciences, University of Dundee, Dow Street, Dundee, DD1 5EH, UK. E-mail: a.pisliakov@dundee.ac.uk

[d] Department of Chemistry, Technische Universität Berlin, 10623, Berlin, Germany. E-mail: n.elghobashi-meinhardt@campus.tu-berlin.de

[e] UCL School of Pharmacy, University College London, 29/39 Brunswick Square, London, WC1N 1AX, UK. E-mail: g.wells@ucl.ac.uk

[f] School of Physics & Astronomy, University of Leeds, Leeds, LS2 9JT, UK. E-mail: s.a.harris@leeds.ac.uk

[g] Université de Paris, CiTCoM, CNRS, F-75006, Paris, France. E-mail: elisa.frezza@parisdescartes.fr

† Electronic supplementary information (ESI) available. See DOI: 10.1039/d1sc02775a

‡ Equal contributions.

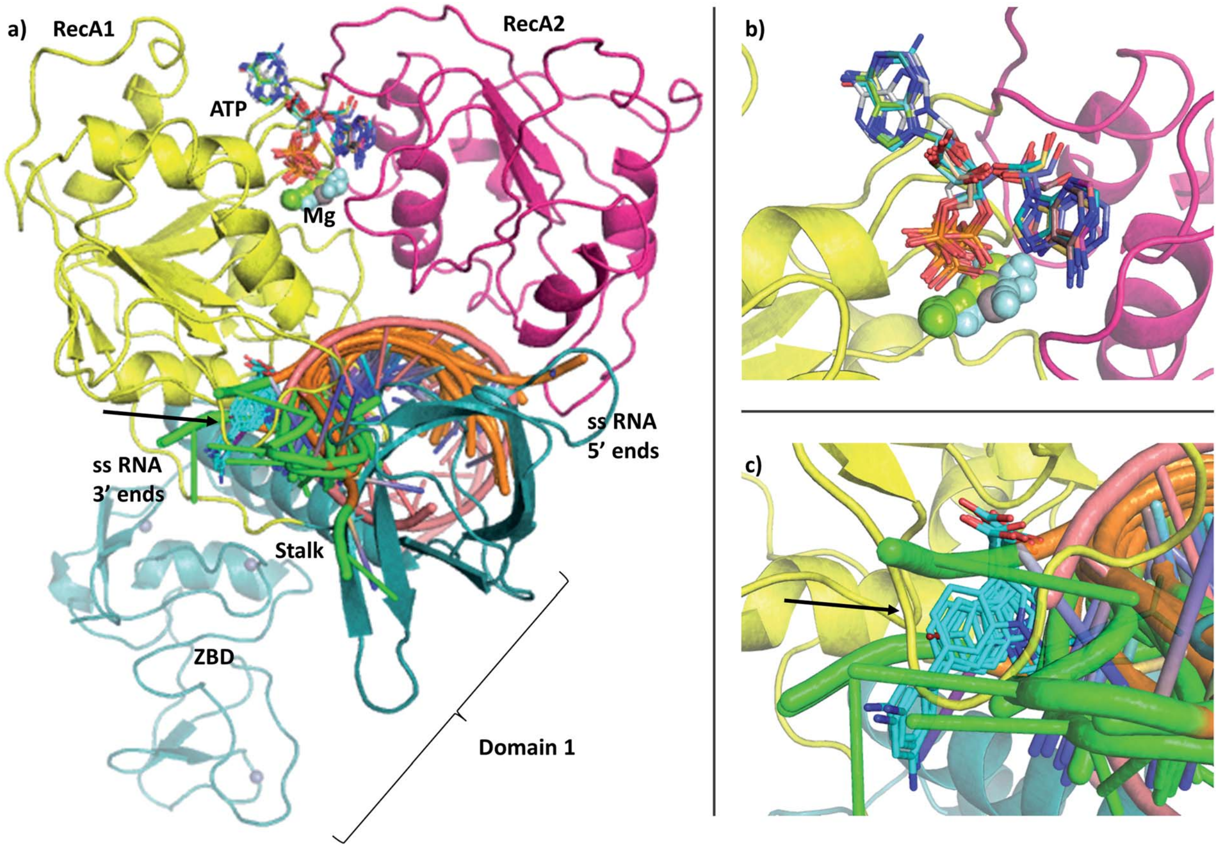

Fig. 1 (a) Our model of the RNA helicase NSP13 of SARS-CoV-2 monomer (cartoon) coloured by three domains: RecA1 (yellow), RecA2 (magenta), and Domain 1 (aquamarine). ATP analogues (sticks) along with Mg (green sphere) and single stranded nucleic acids are depicted from aligned homologous structures (full list of PDB codes are available in Table S1†). 3′ ends of the nucleic acids present the same orientation in all chains (highlighted in green). (b) Position of the ATP analogues (nucleotides in stick and metal ions and compounds in spheres) in homologous structures. (c) Specific helicase inhibitor binding region with allosteric inhibitors displayed in cyan (black arrow).

NTPase activity of the coronaviral helicase, but there are rare examples for selectively hindering the unwinding,[13] or the NTP hydrolysis.[14] Only a few of these efforts were based on or considered structural information. Notably, Hoffmann *et al.* build a homology model and proposed some lead compounds that may interact with the ATP site.[20] Based on the effort on the SARS-CoV helicase, similar approaches can be applied to the highly homologous SARS-CoV-2 helicase.[7]

Coronaviral RNA helicases share a high similarity. 600 out of the 601 residues of the SARS-CoV-2 RNA helicase are identical to those of the SARS-CoV virus, and 70% match that of the MERS-CoV NSP13, demonstrating that these proteins are highly conserved within the *coronaviridae* family. Recently, a set of deposited structures from the PanDDA analysis group (to be published) deposited 51 high resolution crystal structure of the apo SARS-CoV-2 helicases in complex with a library of small molecule fragment analogues.

Helicase structures and models

The first SARS-CoV-2 helicase structure (PDB ID 6zsl) was deposited in July 2020 and the almost identical SARS-CoV helicase structure in 2019 (PDB ID 6jyt),[21] both resolved as crystallographic dimers (Fig. 2a and b). Interestingly, the dimerization interface is different in the two cases, leading to structurally dissimilar complexes. Recent works mainly focusing on the RdRp NSP12,[22,23] which is expressed in the polyprotein sequence just before the helicase, also yielded structures of the replication machinery, including low resolution cryo-EM images of the helicase. In the cryo-EM structure of the RdRp complexed with the RNA helicase (and cofactors NSP7 and NSP8), the two helicase protomers mainly interact with NSP8 (Fig. 2c).[4] The helicase chains were resolved using the apo helicase 6jyt as a template for the cryo-EM density maps and refined using software algorithms.[23] Unfortunately, the 3.7 Å resolution is too low in this structure to resolve the ATP pocket in a catalytically competent conformation.

A more recent high resolution (2.90 Å) cryo-EM work of Lou *et al.*, presents a partial structure of the replication and transcription complex (RTC) (Fig. 2d).[24] The complex includes the RdRp, NSP7, NSP8 and two helicase NSP13 protein copies, with one of the helicases in complex with a single-stranded RNA (ssRNA) fragment. Considering these structure, we conclude that catalytically active form of the helicase is a monomer within the larger RTC complex, and the crystallized homodimeric forms are not the biologically functional unit. Even more recently, a structure (PDB ID 7NNO) was released as a monomer, binding an ATP analogue ANP. The active site agrees with our model, however, this structure does not host ssRNA.

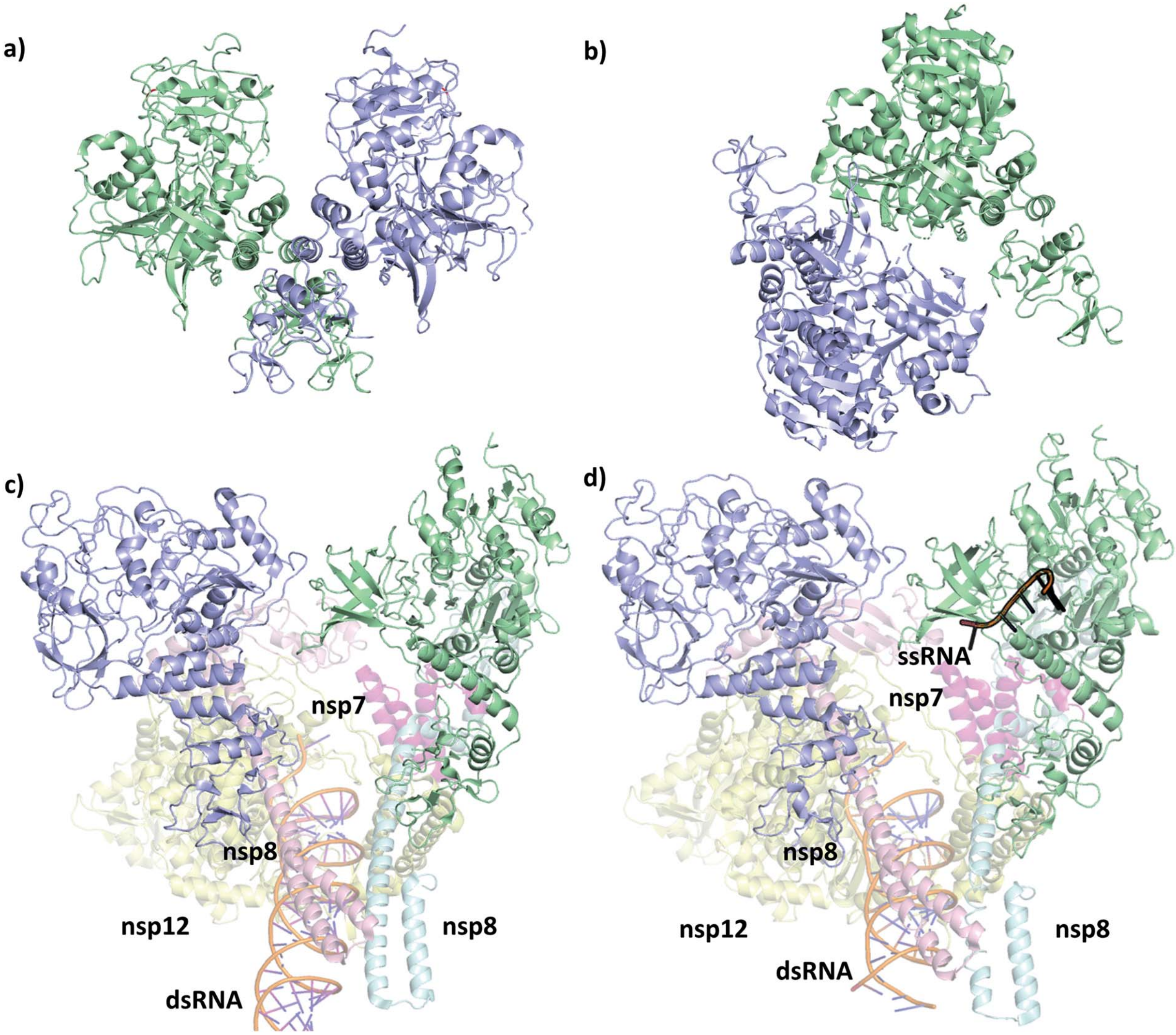

Fig. 2 Structural comparison of the deposited PDB structures of the helicase dimer in SARS-CoV-1 (PDBID: 6jyt, a), SARS-CoV-2 (PDBID: 6zsl, b), SARS-CoV-2 in complex with NSP7 NSP8 and NSP12 (PDBID: 6xez, c) and SARS-CoV-2 with a small fragment of ssRNA bound in complex with NSP7 NSP8 and NSP12 (PDBID: 7cxm, d). The interaction between the two helicase monomers differs depending on the experimental method used to resolve the structures.

Although more attention is directed to targeting the RdRp[22,25–27] or the main protease,[28–32] as these are suggested to be more susceptible for binding an inhibitor,[33–35] the helicase is also subject to modelling and docking studies. Crucially, however, the pharmacophore and docking studies of the helicase start from the homologous SARS-CoV crystal structure (Fig. 2a)[36–40] or the counterpart in SARS-CoV-2 (Fig. 2b),[41] which are both apo, lacking the ATP and the ssRNA from the complex. MD simulations are also available with ATP and ssRNA that used a docking approach to identify promising bound ligands at the ATP binding site.[42]

Here we present a computational model of the SARS-CoV-2 RNA helicase with ATP and ssRNA substrates bound. We performed sequence similarity searches to identify key domains and homologous sequences suggesting structurally important conserved motifs. We also performed structural alignments of available homologous helicase crystal structures to help position the bound RNA and ATP substrates (Fig. 1b). Using both Amber and CHARMM force fields, we carried out long timescale MD simulations of both the apo and the substrate bound states to address the flexibility and the stability of our catalytically competent structures. We analysed the differences in the dynamics between the apo and the holo structures using Principal Component Analysis (PCA). We identified the substrate and allosteric binding pockets and developed an implementation to follow their dynamical behaviour during the MD simulations. We demonstrate novel pockets, including ones that are coronaviridae-specific. Our results will help guide ongoing drug development.

Methods

Homology modelling

Proteins with crystal structures were aligned with MUSTANG for a combined structural-sequence alignment.[43] The apo SARS-

CoV-2 helicase structure was based on PBD ID 6jyt.[21] Missing residues were added and the I570V replacement were carried out in Pymol 2.3.0.[44] The positions of the Mg^{2+} and ATP were determined using the coordinates of PDB ID 2xzo,[45] as a template. Crystallographic water residues were also taken from 2xzo as well as residues around the ATP pocket (loops 284–289 and 534–541, Gln404 and Arg443), except for Arg442 which was modelled based on PBD ID 6jim.[46] The ssRNA was positioned based on 2xzl.[45] The protonation state of titratable residues were estimated by PROPKA 3.0 (Tables S2 and S3†).[47,48]

Molecular dynamics

We performed multiple unbiased MD simulations of the helicase in its apo and holo complex. For a detailed explanation of the methods used to parametrize and run the simulations, please refer to the ESI note 1 and Table S4† for a list of all simulations. We performed the MD using three independent force field setups: (1) CHARMM36 combined with TIP3P water potential (CHARMM), (2) Amber14SB for the protein, Amber ff99 + parmbsc0 + chioL3 for the ssRNA and TIP3P for water (Amber), and (3) Amber14SB protein force field combined with ff99OL3 for the RNA and TIP3P for water. We compared the simulations produced with (1) and (2) as part of the analysis and check the convergency by calculating the RMSD (Fig. S1†).

Pocket analysis

MD trajectories were sampled at 1 ps intervals and stripped of all non-protein residues for pocket analysis. All pockets above the volume of 200 Å^3 were obtained by using pyvol,[49] with default parameters (sphere radius 1.4–3.4). Pocket equivalency across frames were based on Euclidean distance measured from every tenth α carbon of the protein backbone.

Principal component analysis

We used PCA to assess the conformational changes observed in the monomer molecular dynamics simulations.[50] The analysis was restricted to the α-carbon protein atoms to reduce the dimensionality of the dataset,[51] and the protein chain was truncated to limit the contribution of end effects.

The data was grouped by the force-field used (CHARMM/Amber), and whether it was an apo or holo structure. The PCA was performed on these large groups using the scikit-learn library.[52,53] Before the decomposition, each protomer in each frame from each simulation was aligned *via* RMSD minimization to a reference structure from the equilibrated holo model.

Weighted RMSD modes N_i were calculated to show the contribution of each of the m residues to the ith PCA mode using the following equation:[51]

$$N_i = \sqrt{\lambda_i}\begin{pmatrix} \sqrt{x_1^2 + y_1^2 + z_1^2} \\ \vdots \\ \sqrt{x_m^2 + y_m^2 + z_m^2} \end{pmatrix}$$

Where λ_i is the i^{th} eigenvalue, and the vector rows correspond to the coordinates describing the positions of each of the m α-carbon atoms. Component-wise decomposition of this vector gives a quantitative assessment of how much each residue influences the respective PCA mode.

We used dynamic cross-correlation (DCC) map analysis to determine inter-residue displacement correlations, calculated using:[54,55]

$$C_{ij} = \frac{\langle \Delta r_i(t) \Delta r_j(t) \rangle_t}{\sqrt{\langle \|\Delta \boldsymbol{r_i}(t)\|^2 \rangle_t} \sqrt{\langle \|\Delta \boldsymbol{r_j}(t)\|^2 \rangle_t}}$$

Where $\boldsymbol{r_i}(t)$ is the ith atoms coordinates at time t, $\langle x \rangle_t$ denotes the time ensemble average of the quantity x, and $\boldsymbol{\Delta r_i}(t) = \boldsymbol{r_i}(t) - \langle \boldsymbol{r_i}(t)_t \rangle$.

This equation yields a scalar quantity C_{ij} for each pair of atoms, in the range 1 to −1. The closer the value to 1, the more the displacement of atom i is correlated to that of j. Similarly, a negative value indicates an inverse correlation between the two displacements, and a zero value indicates there is no correlation. The maps indicate which residues are displaced together, highlighting groups of residues that move as larger units.

Dynamic weighted histogram analysis method (DHAM)

To calculate the free energy surface for the protein conformational landscape corresponding to key dynamical variables, we constructed a discretized two-dimensional grid to determine Markov State Models (MSMs).[56] The collective variables were extracted along the trajectory, by calculating parameters, including inter-atomic distances, puckering angles, PCA components and pocket volumes. The 2D free energy surfaces are calculated from the first eigenvectors of the MSMs, and provide thermodynamic information on the collective variables used.

Results

Helicase domains and their sequence homology

The single-chain SARS-CoV-2 helicase can be divided into five domains.[21] The sequence starts with Domain 1 (residues 1–260), which features: a Zinc-binding domain (ZBD, residues 1–100), known to facilitate nucleic acid recognition;[57] a Stalk region shaped by 2 contiguous alpha helices (residues 100–150) which functions as an interface connecting the ZBD with Domain 1B (residues 150–260) that interacts with the ssRNA. For simplicity, we refer to domain one as these three combined (Fig. 1, aquamarine cartoon). The rest of the chain is divided into RecA1 and RecA2 domains,[58] which are well characterized in the superfamily 1B-type helicases and interact with ATP at their interface (Fig. 1, yellow and magenta cartoon, respectively).

We have obtained the most homologous 957 sequences and their alignments from the UniProtKB library.[59] The pairwise alignments showed flawed identification of RecA1 (see Fig. S2†), therefore we proceeded to optimize a multiple sequence alignment (MSA).[60] Firstly, we clustered the obtained sequences to avoid overrepresenting highly similar entries.[61] The full list of 796 clusters is available in the ESI note 2.

A set of 52 of these clusters and their representative sequences show similarity across the whole helicase sequence and match at least half of the helicase in the MSA (Fig. 3a, lime region). These represent 96 sequences, all derived from coronaviruses, primarily originate from human and bat viromes (beta and alphacoronaviruses),[62] and infect various hosts in the animal kingdom, including humans. Intriguingly, the next best sequence alignment only matches 107 amino acids; most of these and subsequent aligned regions are specific to the RecA domains and span all types of proteins from various organisms (Fig. 3a blue region).

To evaluate any similarities to Domain 1 only, we also performed a search using only the first 230 residues. This search for sequences that match at least 70 residues resulted in the exact same 96 sequences as before, exclusively belonging to *coronaviridae*. An additional 21 sequences match only shorter segments of the sequence between residues 1–230, corresponding to a 22% sequence identity or below.

The MSA also enables us to see which regions and motifs are conserved in the SARS-CoV-2 helicase sequence. Domain 1 only exhibits a few residues apparently with higher conservation (at positions 87, 124 and 235), however, these are likely to be only random matches, as the corresponding alignments are dominated by non-coronaviral sequences, and not aligned well to the

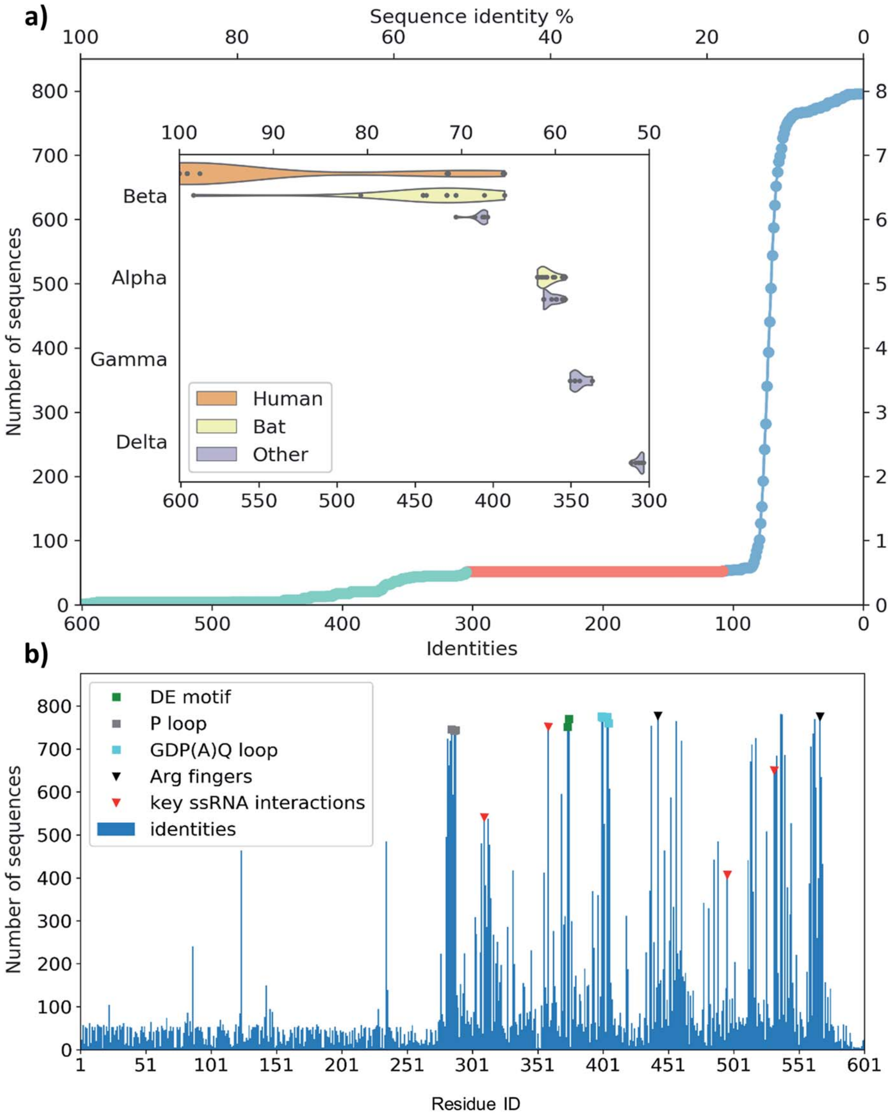

Fig. 3 (a) Distribution of the identity of sequences in multiple sequence alignment compared to the SARS-CoV-2 helicase. There are only members of coronaviridae above 300 matching residues (50%, lime circles, 52 entries). There are no sequences with medium similarity (107–300 similar residues, red circles). The mass of the sequences matches only 107 residues or less (blue circles). The closest relatives (96 sequences represented by 52) are grouped in coronavirus subfamilies (grouped in the *y* axis) with principal hosts highlighted in the inset. (b) Sequence identity of the representative sequences of 796 clusters from UniProtKB aligned to the 601-residue long SARS-CoV-2 RNA helicase. Domain 1 shows similarity only to the close relatives (52 sequences, representing clusters of 96), while the RecA1 and RecA2 domains are more common across ATPase sequences. Key structural motifs are highlighted using symbols (P-loop: grey square, DE motif: green square, arginine fingers: black triangle, ssRNA interactions: red triangles).

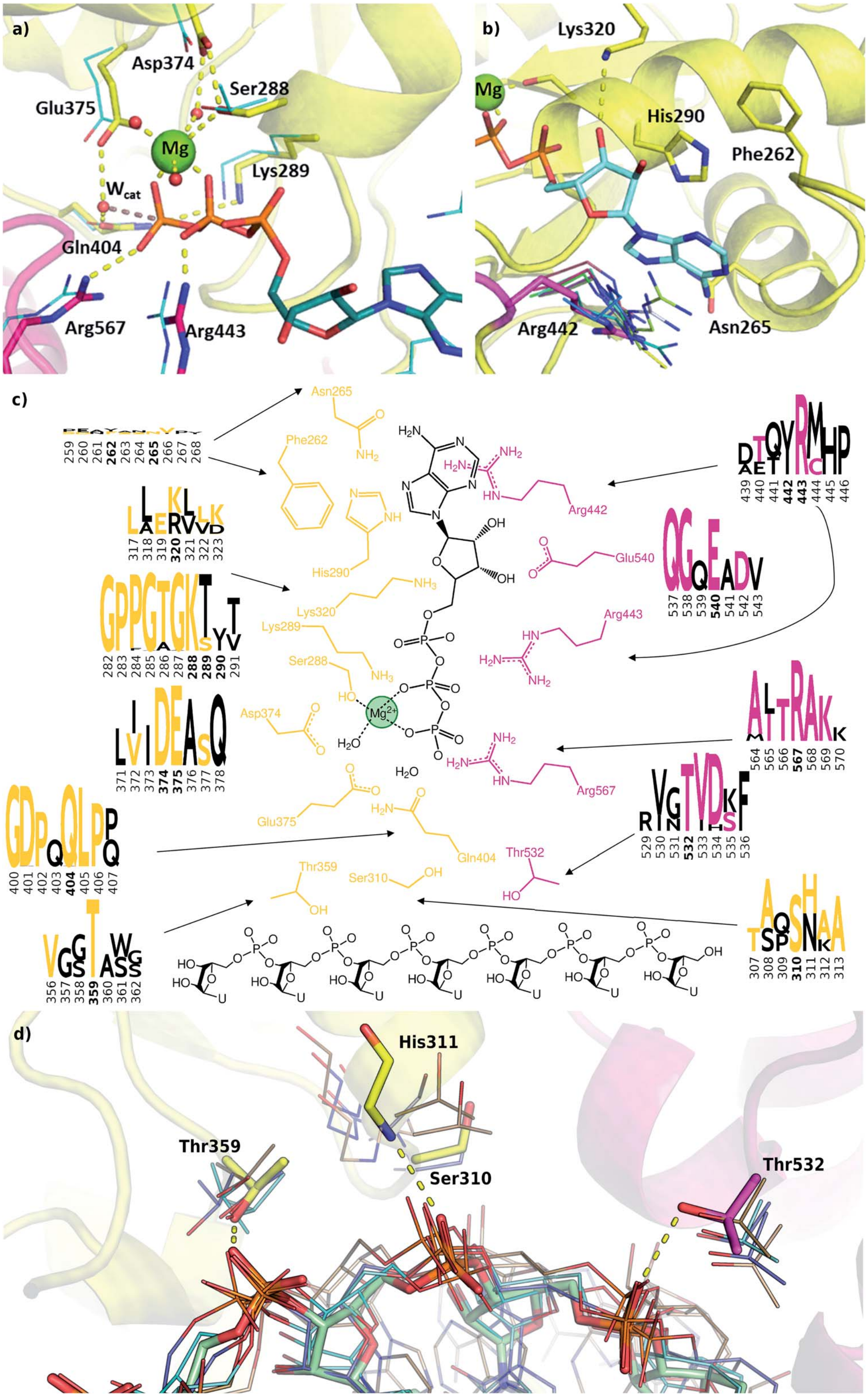

Fig. 4 Modelling and conservation of the ATP and RNA sites. (a) Main protein-substrate interactions of the triphosphate and magnesium ions are compared with alignment for PDB template 2xzo (cyan lines). (b) Nucleotide-binding region focusing on Arg442 (magenta sticks) is aligned with homologous arginine residues (lines, PDB structures 5k8u, 5vhc, 5xdr, 5y4z, 5 y6m, 5y6n, 6adx, 6ady, 6c90 and 6jim). (c) Sequence conservation for RecA1 (orange) and RecA2 (magenta) domains are depicted in logos for each residue and its neighbours (data from Fig. 3b). Coloured letters represent the residues in the SARS-CoV-2 helicase sequence, depicted residue indices are bold in the logos. (d) Structures of the RNA binding region aligned with existing RNA-helicase crystal structures complexed with ssRNA (depicted in lines). RecA1 and RecA2 domains are shown in yellow and magenta, respectively. Key residues (sticks) are labelled, and H-bonds are depicted in yellow dashes.

neighbouring residues of the SARS-CoV-2 helicase sequence. The RecA domains, which are members of the *AAA 30* and *AAA 12* families, are more common in various ATP-binding structures and the MSA indeed sheds light upon important binding motifs such as the P-loop (Fig. 3b, grey squares) the DE motif (Fig. 3b, green squares) and the arginine fingers (Fig. 3b, black triangles). Furthermore, we identified several residues involved in RNA binding (Fig. 3b, red triangles) which are also conserved structurally in homologous PDB entries. Finally, the GDP(A)Q loop at position 400–404 with high consensus in the MSA features a glutamine potentially involved in the proton transfer during the ATP hydrolysis. This motif bridges the γ-phosphate end of the ATP pocket and the RNA binding site; therefore we suggest that it may be involved in coupling the hydrolysis of ATP to the changes induced by the hydrolysis inducing the RNA translocation.

Among crystal structures containing ATP analogues, most helicases have very low sequence similarity to NSP13. The closest homologues are 2xzo, 5mzn, and 6jim with 11.0%, 10.2%, and 8.4% sequence identity, respectively. Despite the low sequence identity, most residues in the ATP binding pocket are conserved. At the same time, the closest human sequence homologue based on our homology search, ZGRF1, a putative RNA helicase, shares only 22% sequence similarity, restricted to the RecA1 and RecA2 domains. This relatively narrow bandwidth of sequence similarity may be advantageous to the design of specific inhibitors against the coronavirus RNA helicases that do not inhibit human proteins.

Structural model of the ATP binding site

We modelled the ATP-bound active site of the SARS-CoV-2 helicase using the 2xzo structure as a template.[45] The essential Mg^{2+} ion cofactor coordinates both the β- and γ-phosphates and a conserved Ser288 (Fig. 4a and c). The active site contains a DE of the DEAD-motif of RNA helicases. The conserved Asp374 H-bonds with Ser288 and one of the Mg-coordinating water molecules, whereas the Glu375 is positioned as the proton acceptor.[20,63,64] The γ-phosphate is stabilized *via* electrostatic interactions and H-bonds with Arg567 Lys289 and Gln404 through a water molecule, which are also found in 95, 56 and 57% of the homologous sequences analysed, respectively. The β-phosphate forms a H-bond with Arg443. Unlike the highly conserved residues recognizing the triphosphate pocket, the environment of the sugar and purine moieties (Fig. 4b and c) shows a greater diversity. The ATP ribose is likely to interact with Glu540 and Lys320 as seen in ten and four homologous PDB structures, respectively. The purine ring is stabilized through multiple π stacking interactions, from one side with Arg442 (a π-cation interaction), in some helicases this interaction is fulfilled with a tyrosine residue; from the other side with His290 and Phe261. Additionally, there is a H-bond between the exocyclic amino group of the purine with Asn265, a residue which is more typically served by a glutamine in similar sequences.

A lack of specificity towards the purine group is likely due to the dual function of the SARS-CoV-2 helicase to aid the 5′-capping of the RNA by the triphosphate hydrolysis of most NTP substrates.[9] Due to these major differences, this area of the nucleotide-binding pocket may be useful in the design of SARS-CoV-2-specific antiviral drugs.

Structural model of the RNA binding site

The most significant changes between the holo and apo structures are related to the binding of the large ssRNA substrate. This substrate binding is more challenging to model, partly due to potential force field inaccuracies, and partly also due to the less specific interactions between the protein and the RNA sidechain that has to accommodate a range of viral sequences for the unwinding and translocation function of the helicase. Despite the relatively large size of the ssRNA substrate, we did not observe large scale domain movements in the holo structure compared to the apo. We observed more localized conformational changes: only the loop of residues 482–487 and the C terminus of RecA2 domain changed considerably compared to the apo structure.

Filtering the related crystal structures those containing nucleic acids (NAs), we noticed that their directionality relative to the ATP pocket is well defined (Fig. 1a and c). Domain 1, being in contact with the sidechain of the ssRNA, does not feature specific motifs, thus allowing different RNA bases to translocate. A long loop transitions into the RecA1 and 2 domains sandwiching the ATP pocket on the side of the ssRNA backbone. This region, equipped with the necessary functionalities to perform the ATP hydrolysis, has a higher degree of conservation along the helicases. Both RecA domains have specific residues that contact ssRNA phosphates, depicted in Fig. 4c and d. Thr359 in RecA1 and Thr532 are identified as the main anchoring points of the two domains. The base between these two threonine residues is coordinated by the backbone NH of His311, an interaction which is conserved in NA containing crystal structures. Ser310 is also reasonably conserved, although not directly involved in ssRNA coordination in this state of the enzyme.

Interestingly, the most conserved motif across the sequences is a GDP(A)Q loop interfacing between the RecA1 and RecA2 domains. This motif features Gln404, a residue which we consider to be important in the coordination of the nucleophilic water in the ATP binding pocket; moreover, it bridges the ATP γ-phosphate and the SH motif discussed earlier. We speculate these moieties play a role in the translocation of the RecA2 unit upon ATP hydrolysis.

MD simulations

ATP binding site. We extracted nine key distances (Fig. 4a and S3†) along the simulations from our unbiased trajectories of the ATP-ssRNA-helicase (holo) complex. There is an overall good agreement between the two force fields for the distribution of these contacts. The hexacoordinate Mg^{2+} shows stable coordination to 3 water molecules, the OH of Ser288 and oxygens from the β- and γ-phosphates of the ATP (Fig. S3a–c†), which is essential for the preorganization of the ATP hydrolysis. Further conserved contacts in the pocket are also maintained during the simulation including the arginine fingers (Fig. S3g and h†), Lys288 (Fig. S3i†) and the DE motif (Fig. S3d and e†) which both takes part in coordinating the Mg^{2+} and the nucleophilic water. The largest deviations between CHARMM and AMBER force fields are observed for the Gln404 and ATP distance (Fig. S3f†). This is a particularly important conserved residue that likely coordinates the attacking water. Using the CHARMM fore field, Gln404 shows greater flexibility deep in the ATP pocket, which might support a role in changing the protein conformation during translocation. Residues participating in the adenosine base coordination are less conserved and form fewer stable contacts.

RNA binding site. From the structural analysis and the homology modelling, we denote two important and well-conserved interactions between the ssRNA and the helicase. Both interactions involve a H-bond between a threonine (Thr359 from RecA1 and Thr532 from RecA2) and a phosphate oxygen on the backbone of the ssRNA. Additionally, another H-bond is made between the central RNA residue and the N of His311, this residue is moderately conserved and the interaction between its backbone and the phosphate oxygen of the RNA is present in several PDB structures. A key residue close to the RNA pocket is Ser310; this residue is conserved (often present as a threonine) and appears to be important for the communication between the ATP pocket and the RNA pocket (Fig. 4c and d).

The stability and dynamics of the RNA have been further analysed by looking at the furanose ring conformation. The definition of the envelope conformers and the puckering angles as descriptors of the conformation are discussed in the ESI Note 3. RNA nucleotides, differently from DNA nucleotides, usually adopt a C3′-endo configuration (usually defined as N), and they become less stable/more reactive when switching to a C2′-endo configuration (defined as S). From a structural comparison with the PDB structure with NA bound to helicase, we can see that most nucleotides present a C2′-end configuration (Fig. S5†). We calculated the puckering value in our MD simulations, as expected most of the nucleotides, present a C3′-endo configuration, relatively stable along with the simulations (Fig. 5 and S4†). Uracil 5, 6 and 7 present a bimodal distribution, showing during multiple trajectories both N and S configuration. Using 2D-DHAM, we calculated the free energy surface by correlating the puckering angle of uracil 6 and the distance between the phosphate oxygen of uracil 6 and the γ-oxygen of Thr532 (Fig. 5a). The reconstruction of the corresponding free energy profile is not possible using the CHARMM trajectories, because the transition between the puckering states is not sampled well. This implies that the CHARMM force field describes the RNA residues more rigidly. We also observe a difference in the orientation of the 3′ terminal residues (Ura7 and Ura8) between force fields, while the rest of the ssRNA behaves similarly and agrees well with the experimental structure (Fig. S6†).

Principal component analysis

To understand the key structural components corresponding to the longer time-scale thermal motions of the protein in the holo (with both ATP and ssRNA bound) and apo forms, we performed principal component analysis. For all simulations analysed, the

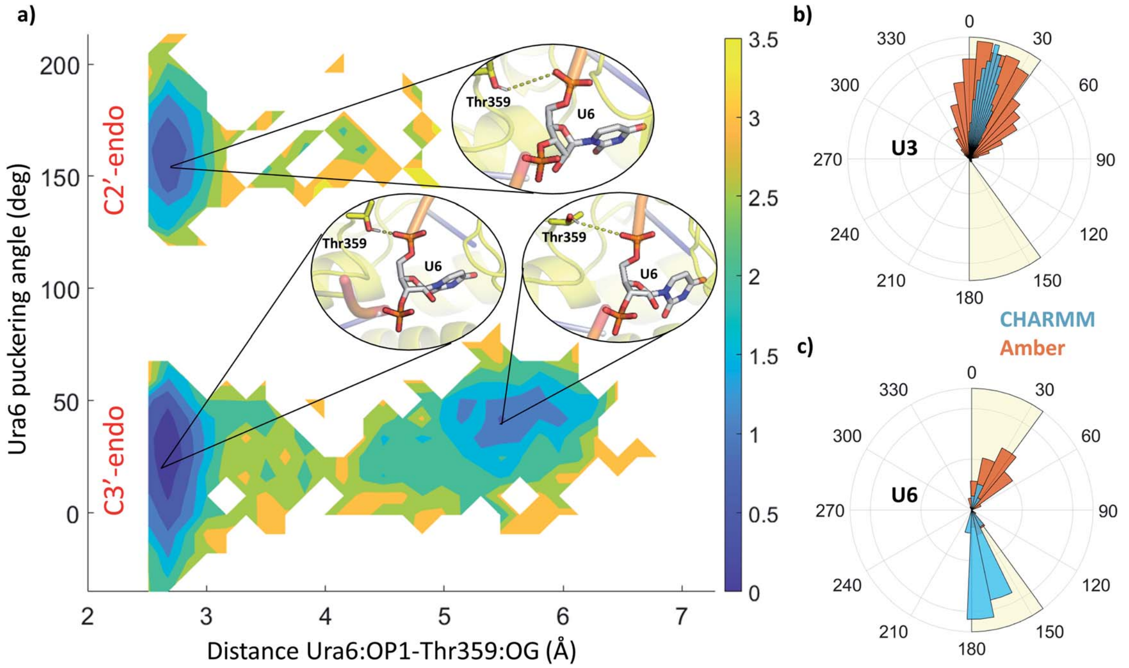

Fig. 5 (a) 2D free energy profile along with the puckering angle and the distance between uracil 6 (OP1) and threonine 359 (OG), from all Amber holo simulations. The colour bar represents the hight of the free energy profile in kcal mol^{-1}. Insets depict the structures in the three local minima, showing uracil 6 in grey sticks and threonine 359 as yellow sticks. Specific distance between the residues is highlighted by yellow dashes. (b and c) Distribution of the puckering angle along the MD simulations using CHARMM (blue) and Amber (orange) force field for uracil 3 (b) and 6 (c).

first four PCA components always accounted for greater than 80% of the observed variance, we therefore focused our analysis on these.

Key observations are shown in Fig. 6, S7 and S8,† which provide a direct comparison between the apo and holo monomer simulations. The weighted RMSD modes for the first four PCA components show the residue displacements captured by each component (Fig. 6a and b and S7†) for the apo and holo simulations, respectively. To reveal the correlated motion within the protein, DCC maps were generated using the first PCA component, which accounts for the largest portion of the overall observed variance (Fig. 6c and e, and S8†).

Increased flexibility in the RecA2 domain is the most clear and significant difference present between the apo and holo data (Fig. 6a and b for Amber and, consistently, in Fig. S7† for CHARMM). Residues around the RecA1 interface and ATP binding site, most notably in the outermost loop from Thr450 to Ala510, are more flexible in the absence of ATP, as is clear from the magnitude of the associated weighted RMSD mode (Fig. 6a and d; label C, red). This observed flexibility is also consistent with the increased experimental beta-factors of this region for the 6jyt and 6zsl structures (ESI note 3 and Fig. S9†). On the other hand, the presence of ATP in the holo simulations stabilizes many of the key residues involved in binding along with the whole of the RecA2 domain (Fig. 6a and b), as it is indicated by the decreased weighted RMSD in the holo simulations and thus smaller contribution to all the PCA components.

The behaviour of the loop from Ile334 to Gln354 is another key difference between holo and apo simulations. This is more prevalent in the Amber simulation PCA components (Fig. 6a and d; label B, purple) and visible as the large cross in the DCC maps (Fig. 6c), which shows its motion differs from the rest of RecA1. Its position between both key RNA binding residues (Thr359 and Thr532) and ATP binding residues (such as Lys320) point towards the substrates providing some tension keeping this loop in place. The stable orientation and position of the

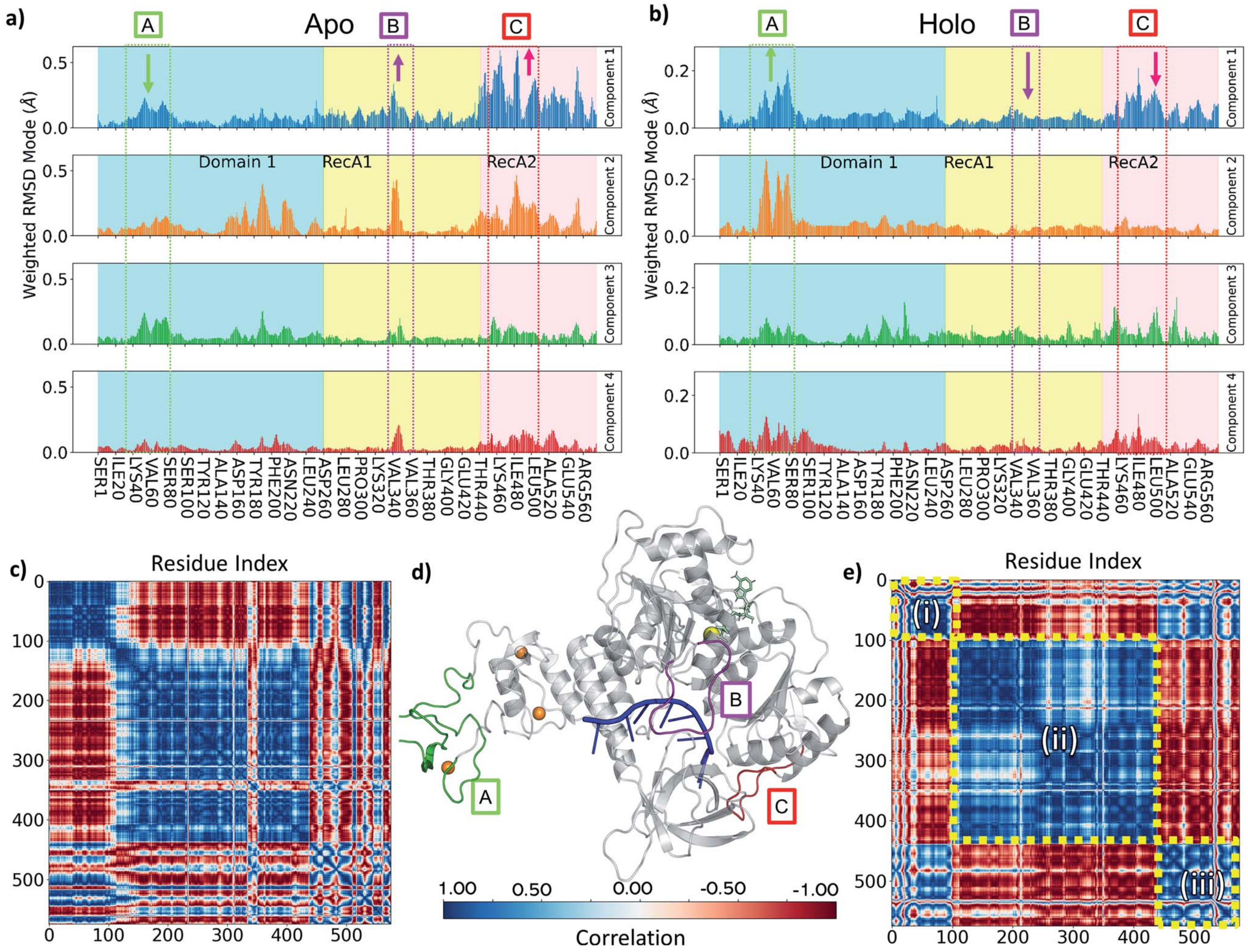

Fig. 6 (a and b) Comparison of PCA Weighted-RMSD scores holo monomer (a) and apo monomer (b) simulations with Amber. (c) DCC map of holo monomer simulation with Amber. (d) Key areas highlighted in the helicase structure (d, grey cartoon) and on panels a and b (text box and arrows indicating relative change in magnitude): (A) Zinc Binding region (green), (B) Domain 1 loop from Ile334 to Gln354 (purple), (C) RecA1/2 interface loop from Thr450 to Ala510 (red). Bound objects are shown and coloured: RNA in blue (centre), ATP in lime (upper right), zinc in orange (left spheres) and magnesium as a yellow sphere. (e) DCC map of apo monomer simulation with Amber for motions described by the first principal component. Colour bar providing correlation scale is shown in the bottom centre.

sidechain Lys345 observed during holo simulations alludes to an interaction with either the nearby α-helix containing Lys320 or the ATP binding site itself.

In the holo simulations, PCA identifies more fluctuations from regions in the ZBD (Fig. 6b and d; label A, green). The holo DCC map (Fig. 6e) highlights the predominant conformational motion in the first PCA component, splitting the helicase into three correlated areas: (i) the ZBD domain, (ii) the rest of Domain 1 along with RecA1, and (iii) RecA2. The structure oscillates with the ZBD and RecA2 regions moving in correlation, opposing the larger central region. All bound simulations showed this global motion (Fig. S10†), displaying an increased flexibility in loops around the ZBD domain not directly involved in zinc-binding and a larger correlation across the motion of the RecA2 domain. However, while the scale of these movements does not indicate a domain level change, it is indicative of the protein flexibility otherwise not easily accessible from *e.g.*, crystallographic data, and disrupting this major motion may provide an aim for future inhibitor design.

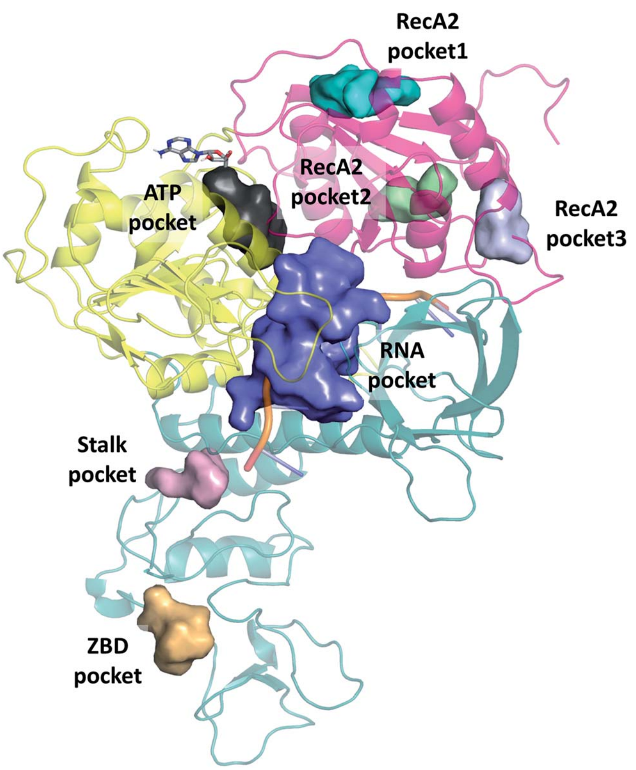

Fig. 7 Overview of binding pockets (coloured surfaces) identified in the holo complex MD trajectories. ATP (grey sticks) and the ssRNA (cartoon) are only show for the sake of orientation.

Pocket analysis

We selected cavities that were consistently present in our MD simulations and tracked the changes in these pocket volumes during the trajectories (Fig. 7). The average volumes and corresponding standard deviations of the pocket sizes in different simulation types are also compared with available experimental structures in Table 1. The distributions of the pocket volumes are depicted in Fig. S9 and S10† for the holo and apo systems, respectively.

Substrate pockets. The ATP pocket is easily identified in the holo trajectories and its volume is ∼600 Å^3. In the apo trajectories the standard deviation increases, indicating less constrained movement between RecA1 and RecA2. The RNA pocket is usually the largest cavity identified and ranges along the interface between the domains, largely overlapping with the RNA binding site. The volume of the RNA pocket increases from 1911 (1472 with CHARMM) to 2415 Å^3 in the apo trajectories compared to the holo structures (Table 1), which can be attributed to the larger freedom in the movement of the domains as the RNA goes along the interface. Additional data on the apo dimer simulations are also available in Table S5.†

The 6jyt crystal structure features a connected ATP-RNA pocket in both of the monomer chains, as is often observed in

Table 1 Occurrence and volume statistics of pockets depicted in Fig. 7 in MD simulations and available PDB structures. The rows in the experimental structures represent the different chains. Note that the minimum volume for any pocket to be defined is 200 Å^3

		Force field	ATP pocket	RNA pocket	RecA2 pocket1	RecA2 pocket2	RecA2 pocket3	Stalk pocket	ZBD pocket
(%)	Holo	CHARMM	96.28	98.60	18.45	22.25	54.57	26.52	16.91
		Amber	98.28	98.18	34.92	18.12	—	16.40	20.26
	Apo	Amber	81.40	99.74	35.06	—	37.00	17.77	7.89
Volume (stdev)/Å^3	Holo	CHARMM	630 (263)	1472 (858)	276 (64)	409 (179)	405 (200)	449 (202)	268 (74)
		Amber	542 (223)	1911 (565)	330 (91)	353 (119)	—	300 (75)	246 (47)
	Apo	Amber	665 (338)	2415 (873)	303 (106)	—	286 (99)	282 (73)	247 (50)
	6jyt	X-ray	4058		334	204	—	—	—
			4427		1282[a]	—	1282[a]	—	—
	6zsl		474 + 267[b]	—	—	—	—	—	—
			613	—	—	—	—	—	—
	6xez	Cryo-EM	534	3698	662	—	—	279	—
			4891		540	—	—	—	—
	7cxm		708	2780	588	369	—	—	—
			782	1566	—	—	—	—	—

[a] Labelled pockets are connected, the combined volume is shown. [b] Made of two separate pockets.

our apo simulations (Fig. S13†). In general, the sizes of the ATP and RNA pockets agree between the crystal structures and our simulations, including a decrease in the RNA pocket size when ssRNA is bound to the structure (holo, 7cxm in Table 1).

Allosteric pockets. Generally, all non-substrate pockets are significantly smaller, and appear less frequently during the trajectories. Among the RecA2 pockets, pocket1 is the most consistent, while both pocket2 and pocket3 depends on the movement of the C-terminus of the chain and they are sensitive to the presence of the natural substrates or the force field (see Table 1 and S5†). Pocket1 is also found in most of the experimental structures by our pocket analysis.

Domain 1 hosts two interesting pockets, the 'Stalk pocket' and 'ZBD pocket'. The Stalk pocket resides between the longer Stalk helix and the ZBD and can be consistently identified in all trajectories, although its average volume varies, probably because it is close to the N-terminus of the chain, which moves relatively freely. More importantly, several bound molecules were identified at this site experimentally in the helicase-small molecule crystal structures deposited in the PDB (5rli, 5rmd and 5rm1). The Stalk pocket is also identified in one of the cryo EM structures (PDB 6xez). The ZBD pocket is found in most trajectories but in the lowest consistency and average volume among all analysed pockets. The residues neighbouring these pockets are detailed Fig. S14†. Overall, both force fields identified the pockets with similar statistical parameters. One difference was in the additional pocket3 observed in RecA2 with CHARMM that is absent from Amber.

Features influencing the pocket volume

We examined the correlation between the volume of the identified pockets and the residue–residue distances in the trajectories (ESI note 5 and Table S6†). This enables us to detect the mechanism and elucidate the precise structural interactions leading to these cavities opening up and closing during the MD simulations. Subsequently, we also calculated the quantitative free energy profiles related to these correlations using DHAM.

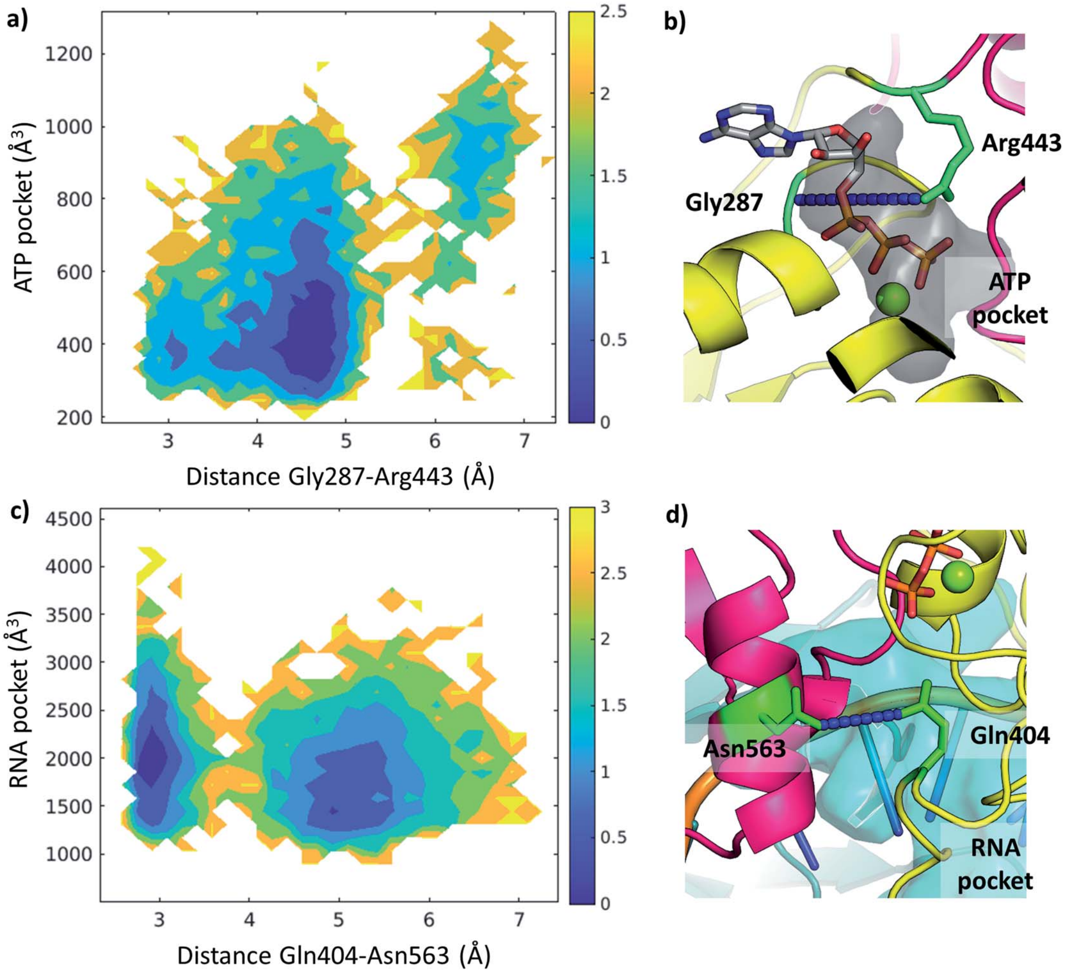

Fig. 8 Free energy profiles depicted along selected coordinates and substrate pocket volumes (holo Amber trajectories). (a) 2D profile along Gly287-Arg443 distance and ATP pocket volume. The color bar represents the height of the free energy profile in kcal mol^{-1}. (b) The ATP pocket (black surface) and the Gly287-Arg443 distance (residues in green, distance in blue) depicted in a representative structure. (c) 2D profile along Gln404-Asn563 distance and RNA pocket volume. The color bar represents the height of the free energy profile in kcal mol^{-1}. (d) The RNA pocket (cyan surface) and the Gln404-Asn563 distance (residues in green, distance in blue) depicted in a representative structure.

The ATP pocket is located at the interface of RecA1 and RecA2, thus it primarily depends on the interactions of the contacting residues. We observed that the opening of the pocket can be described by the distance between Gly287 or the P-loop and the arginine finger 443, resulting in three connected local minima on the potential energy surface (Fig. 8a and b).

On the other hand, the RNA pocket is larger, it has more bordering residues and it can open up in multiple directions. Here, as one of the most prominent directions, we focus on the influence of the contact between Asn563 and Gln404, which is anti-correlated with the size of the RNA pocket. Gln404 is located in one of the most conserved motifs in the sequence (Fig. 3b) and it is connected to the ATP γ-phosphate (Fig. 4c), directly influencing the ATP hydrolysis. The increase of this Asn563-Gln404 distance decreases the optimal RNA pocket size by 500 $Å^3$ down to 1500 $Å^3$ (Fig. 8c and d).

Conclusion

Here, we present structural and sequence comparison studies, as well as molecular dynamics simulations of both the apo and a catalytically relevant computational model of the SARS-CoV-2 NSP13 ATP dependent RNA helicase. The analysis of homologous sequences sheds light upon the specificity of the domain structure of the viral helicase yielding no match over 20% except close relatives from the coronaviridae family.

We performed extensive MD simulations of helicase monomers and a dimer. However, upon analysis of available experimental structures, including the RTC complex, we suggest that the dimer is not the functional unit, and we furthermore focused on the monomer.

To gain key insights into the structure and dynamics of the complete holoenzyme in addition to the experimentally available apo protein, we modelled a fully assembled complex with both the ATP and ssRNA substrates. The structure of the ATP pocket was reconstructed including signature motifs from phosphate binding proteins, such as the DE(AD) of helicases, the P-loop, and arginine fingers. We did not observe large scale domain level motions upon RNA binding. Nevertheless, some conformational changes are required to accommodate the RNA, which, compared to the ATP, does not have so well-defined interactions with the protein to enable multiple sequences to be processed. Moreover, more structural variations and uncertainties for the RNA are also observed in our models during the simulations. Importantly, we identified highly conserved anchoring points in the core of the helicase for polynucleotide binding, which are essential to understand the translocation driving the unwinding activity of NSP13. Our molecular dynamics simulations verified the stability of conserved interactions in our model, as well as improved our initial model to host the nucleic acid. Decomposing the trajectories into principal components highlighted the rigidifying effect of the substrates to the protein structure. The increased stability of our holoenzyme model should be exploited in subsequent docking studies, moving away from the apo structures.

We characterized the volume of the ATP and ssRNA pockets and, importantly, identified additional allosteric binding sites. We assessed the connection between the substrate pockets and key interactions therein, giving insights to the dynamic behaviour of the cavities. Importantly, we found pockets in the highly specific Domain 1 of the helicase which coincides with some experimentally bound substrates. This may provide a good opportunity for specific structure-based inhibitor design.

The comparison of different force fields resulted in small differences only. CHARMM presents a more rigid ssRNA model than Amber, leading to less structural diversity when considering the bound RNA conformation. The ATP binding on the other hand remains robust in all holo simulations with both force fields. The change in the dynamics of the protein upon substrate binding is similar, as well as the qualitative description of allosteric pockets.

Our work provides insight into a key participant of the SARS-CoV-2 viral replication machinery, one of the prominent drug targets. Our structures offer novel starting points for structure-based compound design and screening. The catalytically relevant holo structures are also ideal starting points for subsequent mechanistic studies of the ATPase and the unwinding activity of the helicase. Moreover, elaborating the RNA translocations driven by the identified interactions can reveal other targetable states of the helicase.

Data availability

All simulation data is available at the MolSSI and SlimMD (https://www.ccpbiosim.ac.uk/slimmd) repositories. The pocket analysis code is available at https://github.com/bertadenes/pyvol.

Author contributions

DB: conceptualization, formal analysis (structural and sequence data, MD, pockets), methodology, software; MB: formal analysis (MD), investigation, methodology; SAM: formal analysis (PCA), investigation, methodology; PJB: formal analysis (sequence data); AVP: supervision; NEM: resources, investigation, supervision; GW: investigation, supervision; SAH: resources, investigation, supervision; EF: resources, investigation, supervision; ER: conceptualization, formal analysis, resources, methodology, supervision. All authors contributed to the writing of the paper.

Conflicts of interest

There are no conflicts to declare.

Acknowledgements

ER, DB, MB, and SAM acknowledge funding from ERC Starting grant (project 757850 BioNet). This project made use of time on ARCHER granted *via* the UK High-End Computing Consortium for Biomolecular Simulation, HECBioSim (http://hecbiosim.ac.uk), supported by EPSRC (grant no. EP/R013012/1). We would like to thank GENCI for a generous allocation of computing time at CINES within the action "Project contre le COVID 19" and the project DARI A0080711431. We are also grateful for resourced provided by the Leeds ARC HPC service.

References

1 M. Wang, M. Yan, H. Xu, W. Liang, B. Kan, B. Zheng, H. Chen, H. Zheng, Y. Xu, E. Zhang, H. Wang, J. Ye, G. Li, M. Li, Z. Cui, Y.-F. Liu, R.-T. Guo, X.-N. Liu, L.-H. Zhan, D.-H. Zhou, A. Zhao, R. Hai, D. Yu, Y. Guan and J. Xu, *Emerging Infect. Dis.*, 2005, **11**, 1860–1865.

2 F. Wu, S. Zhao, B. Yu, Y.-M. Chen, W. Wang, Z.-G. Song, Y. Hu, Z.-W. Tao, J.-H. Tian, Y.-Y. Pei, M.-L. Yuan, Y.-L. Zhang, F.-H. Dai, Y. Liu, Q.-M. Wang, J.-J. Zheng, L. Xu, E. C. Holmes and Y.-Z. Zhang, *Nature*, 2020, **579**, 265–269.

3 J. H. Beigel, K. M. Tomashek, L. E. Dodd, A. K. Mehta, B. S. Zingman, A. C. Kalil, E. Hohmann, H. Y. Chu, A. Luetkemeyer, S. Kline, D. Lopez de Castilla, R. W. Finberg, K. Dierberg, V. Tapson, L. Hsieh, T. F. Patterson, R. Paredes, D. A. Sweeney, W. R. Short, G. Touloumi, D. C. Lye, N. Ohmagari, M. Oh, G. M. Ruiz-Palacios, T. Benfield, G. Fätkenheuer, M. G. Kortepeter, R. L. Atmar, C. B. Creech, J. Lundgren, A. G. Babiker, S. Pett, J. D. Neaton, T. H. Burgess, T. Bonnett, M. Green, M. Makowski, A. Osinusi, S. Nayak and H. C. Lane, *N. Engl. J. Med.*, 2020, **383**, 1813–1826.

4 J. Chen, B. Malone, E. Llewellyn, M. Grasso, P. M. M. Shelton, P. D. B. Olinares, K. Maruthi, E. Eng, H. Vatandaslar, B. T. Chait, T. Kapoor, S. A. Darst and E. A. Campbell, *Cell*, 2020, **182**, 1560–1573.

5 Y. Wang, D. Zhang, G. Du, R. Du, J. Zhao, Y. Jin, S. Fu, L. Gao, Z. Cheng, Q. Lu, Y. Hu, G. Luo, K. Wang, Y. Lu, H. Li, S. Wang, S. Ruan, C. Yang, C. Mei, Y. Wang, D. Ding, F. Wu, X. Tang, X. Ye, Y. Ye, B. Liu, J. Yang, W. Yin, A. Wang, G. Fan, F. Zhou, Z. Liu, X. Gu, J. Xu, L. Shang, Y. Zhang, L. Cao, T. Guo, Y. Wan, H. Qin, Y. Jiang, T. Jaki, F. G. Hayden, P. W. Horby, B. Cao and C. Wang, *Lancet*, 2020, **395**, 1569–1578.

6 I. F.-N. Hung, K.-C. Lung, E. Y.-K. Tso, R. Liu, T. W.-H. Chung, M.-Y. Chu, Y.-Y. Ng, J. Lo, J. Chan, A. R. Tam, H.-P. Shum, V. Chan, A. K.-L. Wu, K.-M. Sin, W.-S. Leung, W.-L. Law, D. C. Lung, S. Sin, P. Yeung, C. C.-Y. Yip, R. R. Zhang, A. Y.-F. Fung, E. Y.-W. Yan, K.-H. Leung, J. D. Ip, A. W.-H. Chu, W.-M. Chan, A. C.-K. Ng, R. Lee, K. Fung, A. Yeung, T.-C. Wu, J. W.-M. Chan, W.-W. Yan, W.-M. Chan, J. F.-W. Chan, A. K.-W. Lie, O. T.-Y. Tsang, V. C.-C. Cheng, T.-L. Que, C.-S. Lau, K.-H. Chan, K. K.-W. To and K.-Y. Yuen, *Lancet*, 2020, **395**, 1695–1704.

7 S. Yuan, X. Yin, X. Meng, J. F.-W. Chan, Z.-W. Ye, L. Riva, L. Pache, C. C.-Y. Chan, P.-M. Lai, C. C.-S. Chan, V. K.-M. Poon, A. C.-Y. Lee, N. Matsunaga, Y. Pu, C.-K. Yuen, J. Cao, R. Liang, K. Tang, L. Sheng, Y. Du, W. Xu, C.-Y. Lau, K.-Y. Sit, W.-K. Au, R. Wang, Y.-Y. Zhang, Y.-D. Tang, T. M. Clausen, J. Pihl, J. Oh, K.-H. Sze, A. J. Zhang, H. Chu, K.-H. Kok, D. Wang, X.-H. Cai, J. D. Esko, I. F.-N. Hung, R. A. Li, H. Chen, H. Sun, D.-Y. Jin, R. Sun, S. K. Chanda and K.-Y. Yuen, *Nature*, 2021, **593**, 418–423.

8 J. A. Tanner, R. M. Watt, Y.-B. Chai, L.-Y. Lu, M. C. Lin, J. S. M. Peiris, L. L. M. Poon, H.-F. Kung and J.-D. Huang, *J. Biol. Chem.*, 2003, **278**, 39578–39582.

9 K. A. Ivanov, V. Thiel, J. C. Dobbe, Y. van der Meer, E. J. Snijder and J. Ziebuhr, *J. Virol.*, 2004, **78**, 5619–5632.

10 A. D. Kwong, B. G. Rao and K. T. Jeang, *Nat. Rev. Drug Discovery*, 2005, **4**, 845–853.

11 C. Lee, J. M. Lee, N.-R. Lee, B.-S. Jin, K. J. Jang, D.-E. Kim, Y.-J. Jeong and Y. Chong, *Bioorg. Med. Chem. Lett.*, 2009, **19**, 1636–1638.

12 M. K. Kim, M.-S. Yu, H. R. Park, K. B. Kim, C. Lee, S. Y. Cho, J. Kang, H. Yoon, D.-E. Kim, H. Choo, Y.-J. Jeong and Y. Chong, *Eur. J. Med. Chem.*, 2011, **46**, 5698–5704.

13 A. O. Adedeji, K. Singh, N. E. Calcaterra, M. L. DeDiego, L. Enjuanes, S. Weiss and S. G. Sarafianos, *Antimicrob. Agents Chemother.*, 2012, **56**, 4718–4728.

14 M.-S. Yu, J. Lee, J. M. Lee, Y. Kim, Y.-W. Chin, J.-G. Jee, Y.-S. Keum and Y.-J. Jeong, *Bioorg. Med. Chem. Lett.*, 2012, **22**, 4049–4054.

15 C. Lee, J. M. Lee, N.-R. Lee, D.-E. Kim, Y.-J. Jeong and Y. Chong, *Bioorg. Med. Chem. Lett.*, 2009, **19**, 4538–4541.

16 N. Yang, J. A. Tanner, Z. Wang, J.-D. Huang, B.-J. Zheng, N. Zhu and H. Sun, *Chem. Commun.*, 2007, 4413.

17 A. O. Adedeji, K. Singh, A. Kassim, C. M. Coleman, R. Elliott, S. R. Weiss, M. B. Frieman and S. G. Sarafianos, *Antimicrob. Agents Chemother.*, 2014, **58**, 4894–4898.

18 J. A. Tanner, B.-J. Zheng, J. Zhou, R. M. Watt, J.-Q. Jiang, K.-L. Wong, Y.-P. Lin, L.-Y. Lu, M.-L. He, H.-F. Kung, A. J. Kesel and J.-D. Huang, *Chem. Biol.*, 2005, **12**, 303–311.

19 J.-B. Cho, J.-M. Lee, H.-C. Ahn and Y.-J. Jeong, *J. Microbiol. Biotechnol.*, 2015, **25**, 2007–2010.

20 M. Hoffmann, K. Eitner, M. von Grotthuss, L. Rychlewski, E. Banachowicz, T. Grabarkiewicz, T. Szkoda and A. Kolinski, *J. Comput.-Aided Mol. Des.*, 2006, **20**, 305–319.

21 Z. Jia, L. Yan, Z. Ren, L. Wu, J. Wang, J. Guo, L. Zheng, Z. Ming, L. Zhang, Z. Lou and Z. Rao, *Nucleic Acids Res.*, 2019, **47**, 6538–6550.

22 W. Yin, C. Mao, X. Luan, D.-D. Shen, Q. Shen, H. Su, X. Wang, F. Zhou, W. Zhao, M. Gao, S. Chang, Y.-C. Xie, G. Tian, H.-W. Jiang, S.-C. Tao, J. Shen, Y. Jiang, H. Jiang, Y. Xu, S. Zhang, Y. Zhang and H. E. Xu, *Science*, 2020, **368**, 1499–1504.

23 Q. Peng, R. Peng, B. Yuan, J. Zhao, M. Wang, X. Wang, Q. Wang, Y. Sun, Z. Fan, J. Qi, G. F. Gao and Y. Shi, *Cell Rep.*, 2020, **31**, 107774.

24 L. Yan, Y. Zhang, J. Ge, L. Zheng, Y. Gao, T. Wang, Z. Jia, H. Wang, Y. Huang, M. Li, Q. Wang, Z. Rao and Z. Lou, *Nat. Commun.*, 2020, **11**, 5874.

25 W. Yin, X. Luan, Z. Li, Z. Zhou, Q. Wang, M. Gao, X. Wang, F. Zhou, J. Shi, E. You, M. Liu, Q. Wang, Y. Jiang, H. Jiang, G. Xiao, L. Zhang, X. Yu, S. Zhang and H. Eric Xu, *Nat. Struct. Mol. Biol.*, 2021, **28**, 319–325.

26 K. Kato, T. Honma and K. Fukuzawa, *J. Mol. Graphics Modell.*, 2020, **100**, 107695.

27 R. Pérez-Moraga, J. Forés-Martos, B. Suay-García, J.-L. Duval, A. Falcó and J. Climent, *Pharmaceutics*, 2021, **13**, 488.

28 Z. Jin, X. Du, Y. Xu, Y. Deng, M. Liu, Y. Zhao, B. Zhang, X. Li, L. Zhang, C. Peng, Y. Duan, J. Yu, L. Wang, K. Yang, F. Liu, R. Jiang, X. Yang, T. You, X. Liu, X. Yang, F. Bai, H. Liu, X. Liu, L. W. Guddat, W. Xu, G. Xiao, C. Qin, Z. Shi, H. Jiang, Z. Rao and H. Yang, *Nature*, 2020, **582**, 289–293.

29 K. Świderek and V. Moliner, *Chem. Sci.*, 2020, **11**, 10626–10630.

30 K. Arafet, N. Serrano-Aparicio, A. Lodola, A. J. Mulholland, F. V. González, K. Świderek and V. Moliner, *Chem. Sci.*, 2021, **12**, 1433–1444.

31 C. A. Ramos-Guzmán, J. J. Ruiz-Pernía and I. Tuñón, *Chem. Sci.*, 2021, **12**, 3489–3496.

32 T. Jaffrelot Inizan, F. Célerse, O. Adjoua, D. El Ahdab, L.-H. Jolly, C. Liu, P. Ren, M. Montes, N. Lagarde, L. Lagardère, P. Monmarché and J.-P. Piquemal, *Chem. Sci.*, 2021, **12**, 4889–4907.

33 S.-W. Chan, *Front. Microbiol.*, 2020, **11**, 587944.

34 D. L. McKee, A. Sternberg, U. Stange, S. Laufer and C. Naujokat, *Pharmacol. Res.*, 2020, **157**, 104859.

35 M. R. Freidel and R. S. Armen, *PLoS One*, 2021, **16**, e0246181.

36 R. Pokhrel, P. Chapagain and J. Siltberg-Liberles, *J. Med. Microbiol.*, 2020, **69**, 864–873.

37 M. U. Mirza and M. Froeyen, *J. Pharm. Anal.*, 2020, **10**, 320–328.

38 G. Culletta, M. R. Gulotta, U. Perricone, M. Zappalà, A. M. Almerico and M. Tutone, *Computation*, 2020, **8**, 77.

39 A. B. Gurung, *Gene Reports*, 2020, **21**, 100860.

40 K. Kousar, A. Majeed, F. Yasmin, W. Hussain and N. Rasool, *BioMed Res. Int.*, 2020, **2020**, 6237160.

41 L. Thurakkal, S. Singh, R. Roy, P. Kar, S. Sadhukhan and M. Porel, *Chem. Phys. Lett.*, 2021, **763**, 138193.

42 M. A. White, W. Lin and X. Cheng, *J. Phys. Chem. Lett.*, 2020, **11**, 9144–9151.

43 A. S. Konagurthu, J. C. Whisstock, P. J. Stuckey and A. M. Lesk, *Proteins*, 2006, **64**, 559–574.

44 L. Schrödinger, *The PyMOL Molecular Graphics System, Version 2.3*, 2019.

45 S. Chakrabarti, U. Jayachandran, F. Bonneau, F. Fiorini, C. Basquin, S. Domcke, H. Le Hir and E. Conti, *Mol. Cell*, 2011, **41**, 693–703.

46 Y.-S. Law, A. Utt, Y. B. Tan, J. Zheng, S. Wang, M. W. Chen, P. R. Griffin, A. Merits and D. Luo, *Proc. Natl. Acad. Sci. U. S. A.*, 2019, **116**, 9558–9567.

47 M. H. M. Olsson, C. R. Søndergaard, M. Rostkowski and J. H. Jensen, *J. Chem. Theory Comput.*, 2011, **7**, 525–537.

48 C. R. Søndergaard, M. H. M. Olsson, M. Rostkowski and J. H. Jensen, *J. Chem. Theory Comput.*, 2011, **7**, 2284–2295.

49 R. H. B. Smith, A. C. Dar and A. Schlessinger, *bioRxiv*, 2019, 816702.

50 S. A. M. Stein, A. E. Loccisano, S. M. Firestine and J. D. Evanseck, *Principal Components Analysis: A Review of its Application on Molecular Dynamics Data*, ed. D. C. Spellmeyer, Elsevier, 2006, vol. 2, pp. 233–261.

51 C. C. David and D. J. Jacobs, in *Protein Dynamics: Methods and Protocols*, Springer, 2014, pp. 193–226.

52 F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot and E. Duchesnay, *J. Mach. Learn. Res.*, 2011, **12**, 2825–2830.

53 N. Halko, P. G. Martinsson and J. A. Tropp, *SIAM Rev.*, 2011, **53**, 217–288.

54 J. A. McCammon, *Rep. Prog. Phys.*, 1984, **47**, 1–46.

55 T. Ichiye and M. Karplus, *Proteins: Struct., Funct., Genet.*, 1991, **11**, 205–217.

56 E. Rosta and G. Hummer, *J. Chem. Theory Comput.*, 2015, **11**, 276–285.

57 T. M. T. Hall, *Curr. Opin. Struct. Biol.*, 2005, **15**, 367–373.

58 R. M. Story, I. T. Weber and T. A. Steitz, *Nature*, 1992, **355**, 318–325.

59 A. Bateman, M. Martin, S. Orchard, M. Magrane, R. Agivetova, S. Ahmad, E. Alpi, E. Bowler-Barnett, R. Britto, B. Bursteinas, H. Bye-A-Jee, R. Coetzee, A. Cukura, A. Silva, P. Denny, T. Dogan, T. Ebenezer, J. Fan, L. Castro, P. Garmiri, G. Georghiou, L. Gonzales, E. Hatton-Ellis, A. Hussein, A. Ignatchenko, G. Insana, R. Ishtiaq, P. Jokinen, V. Joshi, D. Jyothi, A. Lock, R. Lopez, A. Luciani, J. Luo, Y. Lussi, A. MacDougall, F. Madeira, M. Mahmoudy, M. Menchi, A. Mishra, K. Moulang, A. Nightingale, C. Oliveira, S. Pundir, G. Qi, S. Raj, D. Rice, M. Lopez, R. Saidi, J. Sampson, T. Sawford, E. Speretta, E. Turner, N. Tyagi, P. Vasudev, V. Volynkin, K. Warner, X. Watkins, R. Zaru, H. Zellner, A. Bridge, S. Poux, N. Redaschi, L. Aimo, G. Argoud-Puy, A. Auchincloss, K. Axelsen, P. Bansal, D. Baratin, M. Blatter, J. Bolleman, E. Boutet, L. Breuza, C. Casals-Casas, E. de Castro, K. Echioukh, E. Coudert, B. Cuche, M. Doche, D. Dornevil, A. Estreicher, M. Famiglietti, M. Feuermann, E. Gasteiger, S. Gehant, V. Gerritsen, A. Gos, N. Gruaz-Gumowski, U. Hinz, C. Hulo, N. Hyka-Nouspikel, F. Jungo, G. Keller, A. Kerhornou, V. Lara, P. Le Mercier, D. Lieberherr, T. Lombardot, X. Martin, P. Masson, A. Morgat, T. Neto, S. Paesano, I. Pedruzzi, S. Pilbout, L. Pourcel, M. Pozzato, M. Pruess, C. Rivoire, C. Sigrist, K. Sonesson, A. Stutz, S. Sundaram, M. Tognolli, L. Verbregue, C. Wu, C. Arighi, L. Arminski, C. Chen, Y. Chen, J. Garavelli, H. Huang, K. Laiho, P. McGarvey, D. Natale, K. Ross, C. Vinayaka, Q. Wang, Y. Wang, L. Yeh and J. Zhang, *Nucleic Acids Res.*, 2021, **49**, D480–D489.

60 K. Katoh and D. M. Standley, *Mol. Biol. Evol.*, 2013, **30**, 772–780.

61 L. Fu, B. Niu, Z. Zhu, S. Wu and W. Li, *Bioinformatics*, 2012, **28**, 3150–3152.

62 R. Lu, X. Zhao, J. Li, P. Niu, B. Yang, H. Wu, W. Wang, H. Song, B. Huang, N. Zhu, Y. Bi, X. Ma, F. Zhan, L. Wang, T. Hu, H. Zhou, Z. Hu, W. Zhou, L. Zhao, J. Chen, Y. Meng, J. Wang, Y. Lin, J. Yuan, Z. Xie, J. Ma, W. J. Liu, D. Wang, W. Xu, E. C. Holmes, G. F. Gao, G. Wu, W. Chen, W. Shi and W. Tan, *Lancet*, 2020, **395**, 565–574.

63 I. Briguglio, S. Piras, P. Corona and A. Carta, *Int. J. Med. Chem.*, 2011, **2011**, 213135.

64 X. Yang, C. Chen, H. Tian, H. Chi, Z. Mu, T. Zhang, K. Yang, Q. Zhao, X. Liu, Z. Wang, X. Ji and H. Yang, *FASEB J.*, 2018, **32**, 5250–5257.