Making the invisible enemy visible

Tristan Croll[1], Kay Diederichs[2], Florens Fischer[3], Cameron Fyfe[4], Yunyun Gao[3,5], Sam Horrell[6], Agnel Praveen Joseph[7], Luise Kandler[3], Oliver Kippes[3], Ferdinand Kirsten[3], Konstantin Müller[3], Kristoper Nolte[3], Alex Payne[8], Matthew G. Reeves[3,5], Jane Richardson[9], Gianluca Santoni[10], Sabrina Stäb[3,5], Dale Tronrud[11], Christopher Williams[9] & Andrea Thorn[3,5]

[1]CIMR, University of Cambridge, UK [2]University of Constance, Germany [3]RVZ, University of Würzburg, Germany [4]Paris, France [5]HARBOR, Unversity of Hamburg, Germany [6]Diamond Lightsource, UK [7]Science and Technology Facilities Council, UK [8]Memorial Sloane Kettering Cancer Center, USA [9]Duke University, USA [10]European Synchrotron Radiation Facility, France [11]Oregon, USA

Abstract

During the COVID-19 pandemic, structural biologists have rushed to solve the structures of the 28 proteins encoded by the SARS-CoV-2 genome in order to understand the viral life cycle and enable structure-based drug design. In addition to the 200 structures from SARS-CoV previously solved, 367 structures covering 16 of the viral proteins have been released in the span of only 6 months.

These structural models serve as basis for research worldwide to understand how the virus hijacks human cells, for structure-based drug design and to aid in the development of vaccines. However, errors often occur in even the most careful structure determination - and are even more common among these structures, which were solved under immense pressure.

From the beginning of the pandemic, the Coronavirus Structural Taskforce has categorized, evaluated and reviewed all of these experimental protein structures in order to help downstream users and original authors. Our website also offers improved models for many key structures, which have been used by Folding@Home, OpenPandemics, the EU JEDI COVID-19 challenge, and others. Here, we describe our work for the first time, give an overview of common problems, and describe a few of these structures that have since acquired better versions in the worldwide Protein Data Bank, either from new data or as depositor re-versions using our suggested changes.

Introduction

SARS-CoV-2, the coronavirus responsible for COVID-19, has a single-stranded RNA genome that encodes 28 proteins. These macromolecules fulfil essential roles in the viral life cycle, enabling SARS-CoV-2 to infect, replicate, and suppress the immune system of its host. For example, the characteristic spikes that protrude from its envelope and allow it to bind to host cells, are a trimer of the surface glycoprotein (Fig. 1). Knowing the atomic structures of these macromolecules is vital for understanding the lifecycle of the virus and helping design specific pharmaceutical compounds that bind and inhibit their functions, with the goal of stopping the cycle of infection.

Since the COVID-19 pandemic hit in the beginning of this year, the structural biology community has swung into action very efficiently and is now strongly engaged to establish the atomic structures of these macromolecules as fast as possible by use of nuclear magnetic resonance (NMR), cryo-electron microscopy (Cryo-EM) and crystallographic methods (*1*). All of these methods require an interpretation of the measured and processed data with a structural model and cannot be fully automated. The resulting structures are made freely and publicly available in the World Wide Protein Data Bank (wwPDB), structural biology's archive of record (*2*). Unfortunately, the fit between model and data is never perfect and errors from measurement, post-processing and modelling are a given. Structures solved in a hurry to address a pressing medical and societal need are even more prone to mistakes. However, as these structures are used for biological interpretation, small errors can have severe

consequences – in particular, in structure-based drug discovery, structural bioinformatics, and computational chemistry; a current focus of SARS-CoV-2 research across the world. As of the writing of this publication, 524 macromolecular structures from SARS-CoV and SARS-CoV-2 have been deposited, covering parts of 16 of the 28 proteins. In this time of crisis, it is therefore vital to ensure that the structural data made available to the wider research community are the best they can be in every regard.

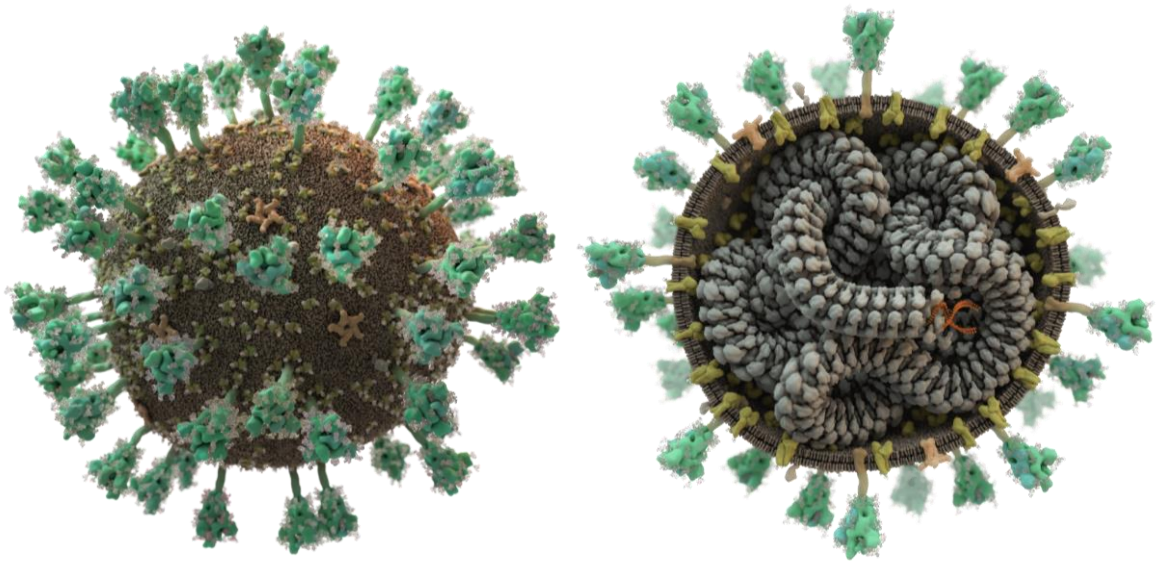

Fig. 1. SARS-CoV-2 displays spike proteins (green) on its surface that recognise and bind to host cells; its lipid bilayer membrane also contains additional embedded membrane and envelope proteins (yellow and beige). The single-stranded RNA (orange) is intertwined in a helical fashion with the nucleocapsid (grey). This figure, however, shows only the transport form of the virus: once a cell is infected, additional viral proteins encoded by the viral RNA are produced that hijack the host cell in order to produce new virus particles. (Picture: Thomas Splettstößer /scistyle.com)

Pushing the methods to the limit

The wwPDB (*3*) is an invaluable tool, but once released, a structure in the databank can only be re-versioned by the original depositors, who, after any associated papers are published, may have little or no motivation to correct their structures. Third parties may only deposit new models based on others' deposited data under new accession IDs when accompanied by peer reviewed publications. In such cases, there is no explicit link from the original entry to the new one. Importantly, 99% of structure downloads from the PDB are not by experimentalists, but scientists from other fields (*2*). As a consequence, errors can lead to a large waste of resources and time by those making use of the data obtained from the PDB and may even be misinterpreted as biologically and pharmaceutically relevant information.

We, the authors of this manuscript, develop computational methods for the solution of macromolecular structures. By being expert users of our own software tools, we were well placed to help in this unprecedented situation. This is why, since the first signs of a possible pandemic arose at the beginning of this year, we joined forces to assess and, where necessary, improve upon the published macromolecular structures from SARS-CoV and SARS-CoV-2. In cases where we believe we have significantly improved the macromolecular models, we offered them back to the original authors and the scientific community. Raw data are not deposited in the wwPDB and are not mandatory for publication; as a consequence, they are difficult to obtain. Their absence in the public record is detrimental for re-analysis and validation, and the development of new methods. In an effort to make the most out of the

experimental results, we invited authors to send us their raw experimental data, which are key to validation of the entire structure determination process from start-to-finish. We thus offer to help authors deposit these data in a non-PDB public repository if the authors wished to do so. After we began our validation efforts of macromolecular structures from SARS-CoV and SARS-CoV-2, we were approached by our colleagues in *in-silico* drug screening, Folding@Home (*4*, *5*), OpenPandemics (*6*), and the EU Joint European Disruptive Initiative (JEDI) (*7*). These initiatives needed the best structures they could get for studies of the virus and had already lost much computing time and resources to suboptimal structure solutions.

Automatic evaluation

All macromolecular structures from SARS-CoV and SARS-CoV-2 in the PDB are downloaded into our repository and assessed automatically in the first 24 hours after release. For crystallographic and Cryo-EM structures, we check the quality of the deposited merged data, and how well the model fits these data. An automatic evaluation of NMR data is forthcoming. Then, all structures are checked according to chemical prior knowledge.

Evaluation specific to crystallographic data and structure solutions

As crystal structures make up 79.0% of our data, these are evaluated most thoroughly. Crystal diffraction can, for example, stem from more than one crystal lattice (twinning), be contaminated by ice crystal diffraction (ice rings) or be incomplete due to radiation damage or suboptimal measurement strategy. These issues cannot be resolved after data collection, but treating data accordingly can yield a better structural model. Deducing such problems from the deposited structure factors (mandatory in wwPDB) can be difficult; raw data allow a much more complete analysis of the experiment.

Another source of errors is data processing (integration and scaling), which nowadays is often done automatically. Assuming the wrong crystal lattice symmetry or including, for example, diffraction spots obscured by the beam stop, can lead to lower quality or even unsolvable structures. If raw data are available, data can be re-processed and these problems can be resolved manually.

To analyse crystallographic data for twinning, completeness and overall diffraction quality, we used phenix.xtriage (*8*); furthermore we ran AUSPEX (*9*) which automatically identifies ice rings and produces plots from which several other pathologies, like a “bad” beam stop mask, can be recognized quickly.

The completeness of most datasets is satisfying, with only 7 out of 415 datasets below 80%. All the datasets have an acceptable strength with intensity/sigma(intensity) above 3. Ice rings were detected in 61 datasets and problems with the beam stop masking in 46; 49 crystal structures were indicated as potentially resulting from twinned crystals.

A general indication of how well the atomic model fits the measurement data can be obtained by comparing the deposited R-factors to results from PDB-REDO (*10*) (including Whatcheck (*11*)) to determine the overall density fit as well as many other diagnostics. While the deposited structures are often improved by PDB-REDO, they need to be checked and should not be viewed as “more correct” purely on basis of a lower R value. In addition to this, a high R value does not indicate a single type of error and hence should be used with caution. Only two structures in the repository present an alarmingly high R_{free} value above 35%, although problems can be found in other structures by looking in more detail. PDB-REDO improves the R_{free} for most of the structures, and the only cases where we found a huge degradation pointed to major issues with the PDB entry; this was especially true for older SARS-CoV structures.

Evaluation specific to structures from single-particle Cryo-EM

Cryo-EM structures make up 15.0% of our data. As with crystallographic structures, raw data are not available from the wwPDB, but the three-dimensional map reconstructed from the microscopic single particle images is deposited, allowing the calculation of the fit between model and map in the form of a Fourier Shell Correlation (FSC). The model-map FSC is plotted as a curve, which estimates agreement between features resolvable at different resolutions. For a well-fitted model, a model-map FSC of 0.5 roughly corresponds to the cryo-EM map resolution (which is determined as where the FSC between two half-maps drops below 0.143). To calculate FSCs, we use the CCP-EM (*12*) model validation task which utilizes REFMAC5(*13*) and calculates real-space Cross-Correlation Coefficient (CCC), Mutual Information (MI) and Segment Manders' Overlap Coefficient (SMOC) (*14*). While MI is a single value score to evaluate how well model and map agree, the SMOC score evaluates the fit of each modelled residue individually and can help to find regions where errors occur in the model in relation to the map. Z-scores highlight residues with a low score relative to their neighbours and point to potential misfits.
Out of 81 structures, 6 structures had an average model-FSC below 0.4 and seven have a MI score below 0.4, indicating a bad overall agreement between map and model and potential for further improvement. The SMOC score indicates for twelve structures that more than 5% of the residues fit poorly with the map, while the other 85% of structures had a relatively good density fit. However, most modelling errors could only be corrected manually (see below).
In addition to this validation, we run Haruspex (*15*), a neural network to annotate reconstruction maps to evaluate which secondary structures can be recognized automatically in the map.

Evaluation of the structural model based on prior knowledge

Molecular geometry is constrained by the nature of its chemical bonds and steric hindrance between the atoms. In order to evaluate the model quality with respect to chemical prior knowledge we run MolProbity (*16*, *17*), which checks covalent geometry, conformational parameters of protein and RNA and steric clashes. However, it is unfortunately possible to use some of these traditional indicators of model quality as additional restraints during refinement, which invalidates them to a certain degree – we therefore also used the MolProbity CaBLAM score (*18*), which can pinpoint local errors at 3-4 Å resolution even if traditional criteria have been used as restraints. CaBLAM scores higher than 2% outliers indicated that 163 of the structures have many incorrect backbone conformations.
During the crisis the MolProbity webservice has been pushed to the limit of its capacity, as many different drug developers have screened the very same coronavirus structures many times. We have developed a bespoke MolProbity pipeline to make these results available online and to decrease the workload on the webservice. In addition to this, the sequence of each structure is also aligned and checked against the known genome to highlight misidentified residues.

Online availability and updates

Every Wednesday, after the new PDB structures have been released, an automatic pipeline runs to organize the new structures according to the genetic information and then to assess the quality of models and the experimental results. These results, along with the original structures, are immediately available from our online repository which is accessible via our website insidecorona.net. To facilitate access and to get an overview of structures, we supply an SQL database of key statistics and quality indicators along with the results.

Manual evaluation

As a community, for decades, we have aspired to automate structural biology as much as possible. However, neither structure solution nor validation have been fully automated due to the complexity of interpreting low-quality maps that have poor fit between experimental data and structural models. This

task requires detailed knowledge of macromolecular/small molecule structure and chemical interactions. Even with state-of-the-art automatic methods at hand, experienced human inspection residue-by-residue remains the best way to judge the quality of a structure, highlighting the continuing need for expert structure solvers. Given the flood of new SARS-CoV-2 structures, resources have not permitted us to check all structures manually. Therefore, we have selected representative structures. Certain errors were surprisingly common, such as peptide bond flips, rotamer outliers and mis-identification of small molecules, such as water as magnesium, chloride as zinc, and a multi-zinc site modelled as poly ethylene glycols. Zinc plays an important role in many viral infections, and is coordinated by many of the SARS-CoV-2 protein structures. We also found a large number of Cys-Zn sites being mismodelled, with the zinc ion missing or pushed out of density, and/or erroneous disulphide bonds between the coordinating cysteine residues. Many coronavirus proteins are linked on certain asparagine residues ("N-linked") to carbohydrate chains called glycans. Their exact composition depends on the host cell in question, and their main function is to deter the host immune system. In some structures, where the sample was produced in eukaryotic cells, the "stem" sugars of these N-linked glycans were evident in the map. However, in many cases these sugars were flipped approximately 180 degrees from their correct orientation around the N-glycosidic bond.

Out of the structures we checked manually, we were able to significantly improve 31 which are available from insidecorona.net. In the following, we will give two examples:

Example 1: Papain-like protease

Once SARS-CoV-2 infects a cell, the first protein produced is a long polypeptide chain which is cleaved into 16 functional proteins, the non-structural proteins (NSPs) (*19*). These are essential for the production of new viruses in the host cell.

NSP3 is a large protein molecule by any measure, 1945 amino acid residues in total. Its 15 segments have a variety of functions, among them the papain-like protease domain which cleaves the first five NSPs from the polypeptide chain (*19*). Without cleavage of the polyprotein, the virus cannot replicate and infection is halted. Hence, the papain-like protease domain represents an important potential drug target (*20*). The first SARS-CoV-2 structure of this domain was PDB 6W9C (released 1st April 2020). It was immediately used as the basis for structure-based drug design around the world. The resolution was 2.7 Å and R_{work}/R_{free} were 23.9% / 30.9%. However, the overall completeness of the measured data was only 57.1%. Why were just over half of all reflections that could have been measured at this resolution recorded?

The raw data for this entry are available from proteindiffraction.org (*21*). They revealed that the diffraction data had been measured with a very high X-ray intensity, which led to a swift deterioration in diffracting power due to radiation damage – something which could not have been learned from the data deposited in the PDB, underlining the importance of the availability of raw data. Typically, crystallographers aim for a dose of 5 MGy or lower. Here, we estimated a diffraction weighted dose of 5.5 MGy with RADDOSE-3D, with a maximum dose value of 21 MGy, which likely completely destroyed parts of the crystal (*22*). We based this calculation on assumption of a 100 x 100 x 10 μm^3 plates, given that the crystals were described as such. In addition, the measurement covered 30° sample rotation followed by an additional 60° rotation starting from the same angular position, covering the first 30° twice, further increasing the dose while recording little additional information. The angular range per image, 0.5°, was also surprisingly wide for the high-throughput pixel detector used (a Dectris Pilatus3 6M), and the diffraction was highly anisotropic.

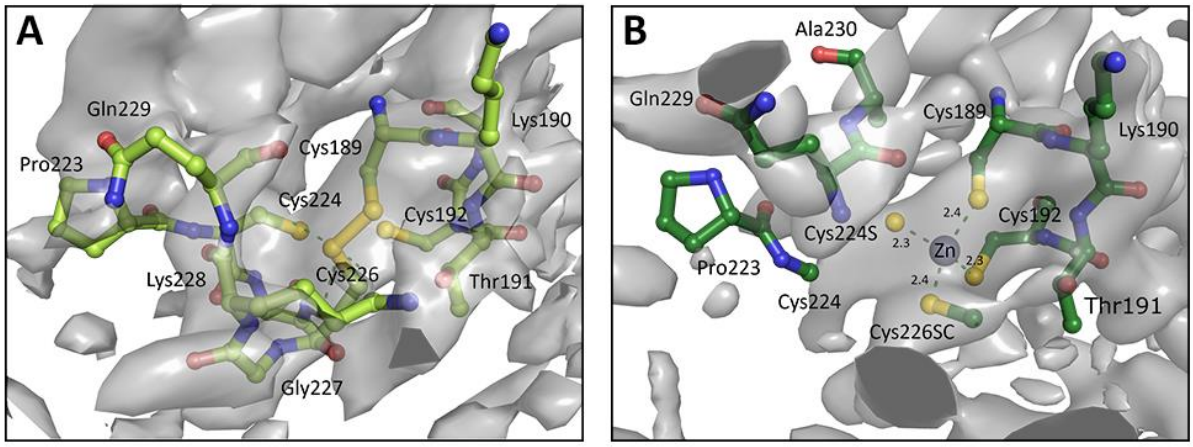

Fig. 2. A. *Chain A of zinc finger from PDB entry 6W9C as deposited, with Cys189 and Cys226 forming a disulphide bond instead of a Zn binding site.* ***B.*** *Re-modelled structure with zinc binding site, utilising 3-fold NCS, prior knowledge about coordination chemistry, and increased geometry weights to improve the map. Electron density is displayed as an isosurface contoured to 1σ.*

We re-processed the images using XDS (*23*), omitting the final 10° of the first sweep and final 20° of the second where the radiation damage affected the data too severely. An elliptical resolution cut-off was applied with Staraniso (*24*) to account for anisotropy. Careful manual intervention could improve the resolution to 2.6 Å with better data quality overall. The revised ellipsoidal completeness was 44.5%.

The structure has 3-fold non-crystallographic rotational symmetry; with three monomers coordinating a central zinc ion within the asymmetric unit. The second, functionally important, zinc ion is far removed from the three-fold rotation axis and coordinated by four cysteines. This zinc finger domain is essential for activity, in addition to the papain-like cysteine-histidine-aspartate catalytic triad (*25*), but it is poorly resolved and incompletely and differently modelled in each of the three monomers of this structure. This disorder may be the result of radiation damage. Only two of the three zinc sites were modelled, and here, the bond lengths between each of the four sulphur atoms and the zinc varied from 2.4 Å to 2.7 Å, and the $C_{ß}$-SG-Zn angles between 70° and 132°. The third site had no zinc (Fig. 2), instead being modelled as a disulphide bond. Prior knowledge about coordination chemistry dictates that the bond lengths between Cys and Zn should all be approximately 2.3 Å and the angles about 107°. Adding zincs to all sites and restraining the bond lengths and angles to these expected values, adding non-crystallographic symmetry restraints (requiring the 3 copies to look similar), an overall higher weighting of ideal geometry, and the reassessment of side chains and water molecules improved the electron density maps and lowered the R values to 20.2%/25.4% at 2.6 Å resolution.

This example shows the importance of optimised data collection strategy, data processing and model building, the quality of which is interconnected. In this case, even though the data were radiation damaged, by adjusting the data processing to take this into account, and by modifying the model refinement to include stronger restraints and to take full advantage of the non-crystallographic symmetry, this structure could be drastically improved. A new structure of the C111S mutant of the same protein (PDB CODE 6WRH) came out a month later, in which the zinc site was clearly resolved. By this time, however, the structure had already been widely used as a target in in-silico drug design: for example, 20% of participants in the EU JEDI COVID-19 challenge have used this structure to design potential drugs. The availability of a better structure a month earlier would not only have increased their chances of success but also saved much computing and man hours in computer aided drug development.

Example 2: RNA polymerase complex

When SARS-CoV-2 replicates, its single-stranded RNA genome needs to be copied. This is achieved by a macromolecular complex of RNA-dependent RNA polymerase (NSP12, RdRp), NSP7 and NSP8 (*26*). Coronaviruses, including SARS-CoV-2, have some of the largest genomes among RNA viruses (approximately 30 kilobases), suggesting their polymerase complexes possess proof-reading functionality. This sets coronaviruses apart from other RNA viruses (*26*).

The first structure of SARS-CoV RNA polymerase (PDB entry 6NUS) was solved in 2019 by Cryo-EM (*27*), before the pandemic began. In this structure, a loop close to the C-terminus (residues 892-906) was not resolved in the reconstruction map and hence not modelled. Following this loop, the polymerase has an irregular helix followed by a flexible tail. Density for this helix was poorly resolved – coupled with its short length and the lack of any information from the preceding and following loops, this led to difficulty in assigning register (the identity of the amino acid at each site). Nevertheless, the overall validation statistics (clashscore, Ramachandran outliers, sidechain outliers) provided by the wwPDB for this model appeared exceptionally good. We inspected one of the first available structures of the SARS-CoV-2 RNA polymerase complex (7BTF) using ISOLDE (*28*), a program used for interactively visualising and remodelling proteins in their experimental density. The higher resolution at this C-terminal tail of the structure made it clear that the C-terminal helix was severely incorrect, with the assigned sequence being nine residues upstream of the correct residues for this site (see Fig.3). This error was present in all the structures of this complex from both SARS-CoV and SARS-CoV-2, presumably propagated due to each subsequent structure using the previous models as the starting point for their modelling, as is standard practice.

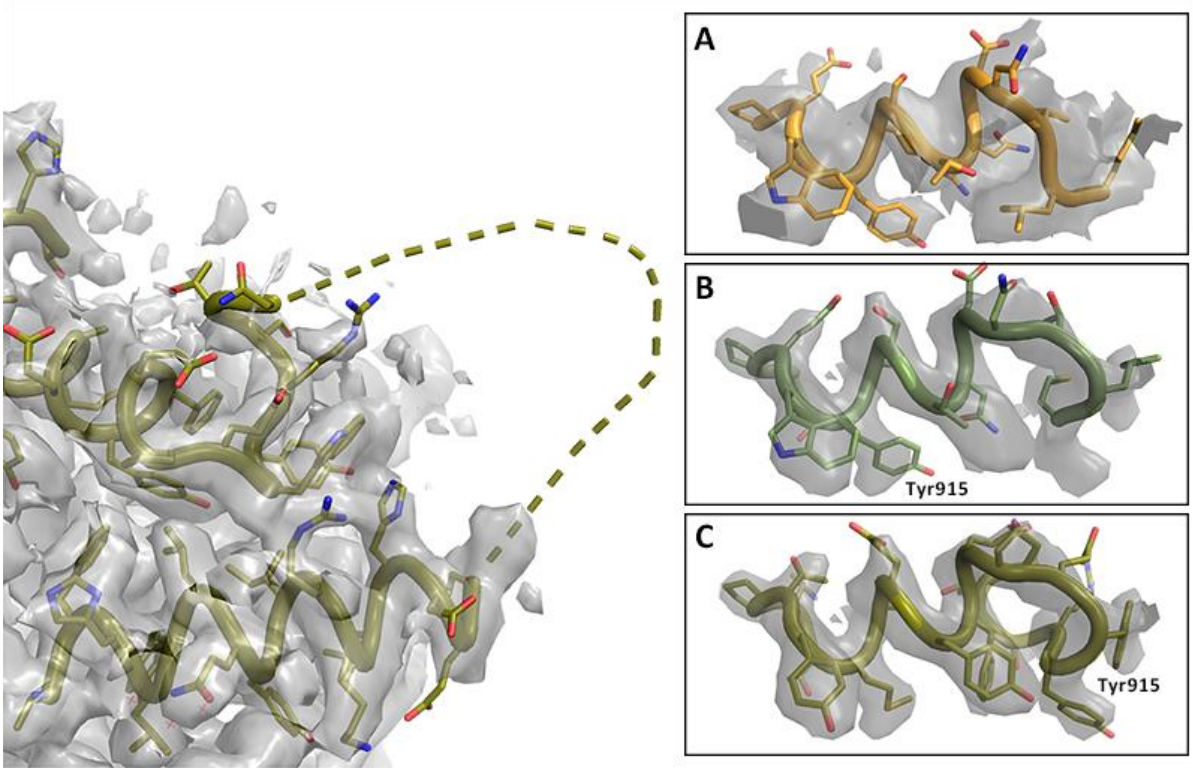

Fig. 3. *Registry shift in C-terminus of RNA Polymerase.* ***Left:*** *Overview with missing loop shown as dashed line (PDB entry 7BV2); map at 2.4σ.* ***Right:*** *Details of C-terminal helix at 5σ.* ***A.*** *Lower resolution map and model PDB 6NUS. Judging the side chain fit is difficult.* ***B.*** *Higher resolution map and model 7BV2 as deposited; the side chain fit is suboptimal.* ***C.*** *Amended 7BV2 structure; the side chains now fit the density. The register shift is indicated by Tyr915.*

For each affected RdRp structure we immediately contacted the original authors. 7 of the 9 SARS-CoV-2 RNA polymerase complexes in the wwPDB now have the corrected sequence alignment at the C-terminus, and those also include many of our other changes described below. These PDB re-versioned corrections allow modelling efforts for drugs against SARS-CoV-2 RNA polymerase to start from a much better model. Notably, the authors of a later cryo-EM structure of the RNA polymerase/RNA complex (PDB entry 6YYT) used one of our corrected models as the starting point for their new model (*29*). The structure of SARS-CoV-2 RNA polymerase in pre-translocation state and bound to template-primer RNA and Remdesivir (PDB entry 7BV2) (*30*) represents a useful basis to investigate the inhibitory effect of Remdesivir and the rational design of other nucleoside triphosphate (NTP) analogues (*31*). However, we found that this structure has some issues, which may provide misleading information to people who are conducting such studies. Apart from the register shift described above, there are three magnesium ions modelled in the active site, a number which is contradictory to our common knowledge of this class of proteins. Magnesium ions play an essential role in catalysis in RNA polymerase (binding the incoming NTP, positioning NTP for incorporation and stabilizing the leaving group after catalysis) (*30*, *32*). One of the magnesium ions is shown coordinated by a pyrophosphate, which implies that the pyrophosphate ion release in SARS-CoV-2 RdRp is relatively slow and may even couple with the translocation (*33*). However, all three magnesium ions as well as the pyrophosphate are poorly supported by the map reconstructed from the experimental data or by local geometry If these ions were included as fixed components of the binding site, this may have severely impacted *in-silico* docking and drug design studies. In addition to the above, we corrected the conformations of three RNA residues close to the Remdesivir site including an adenosine base (T18) modelled “backwards”, fixed “backwards” peptides flagged by CaBLAM, added several residues and water molecules with good density and geometry, and corrected two proline residues that had been erroneously flipped from *cis* to *trans*. Our remodelled structure is offering a valuable structural basis for future studies, such as *in-silico* docking and drug design targeting at SARS-CoV-2 RdRp (*34*), as well as for computational modelling or simulations to investigate the molecular mechanism of viral replication (*31*, *35*, *36*).

Supplying context

Many specialists in structural biology and *in silico* design are now tackling SARS-CoV-2 research, but may not be familiar with the wider body of coronavirus research. In addition to improved models and evaluation results, we also supply context on insidecorona.net. This covers literature reviews centred on the structural aspects of the viral life cycle, host interaction partners, illustrations, and evaluation criteria for selecting the best starting models for *in silico* projects. Furthermore, we added entries about the SARS-CoV-2 proteins to Proteopedia (17) and MolSSI (*37*, *38*), as well as 3D-Bionotes (18) deep-link into our data base. Finally, as SARS-CoV-2 has had an unprecedented impact on the world at large, we have also tried to make our, and others’, research on the topic accessible to the general public. This has included a number of posts on our homepage aimed at non-scientists and live streaming the reprocessing of data on Twitch, as well as the design, production, and public release of an accurate 3D printed model of SARS-CoV-2 based on deposited structures for use as a prop for outreach activities.

Summary

In the last five months, we have done a weekly automatic post-analysis as well as a manual re-processing and re-modelling of representative structures from each of the 16 structurally known macromolecules of SARS-CoV or SARS-CoV-2. In this global crisis, where the community aims to get structures out as fast as possible, we aim to ensure that structure interpretations available to downstream users are as solid as possible. We provide these results as a free resource to the community in order to aid the hunt for a vaccine or anti-viral treatment. Our results are constantly updated and can be found online at **insidecorona.net.** New contributors to this effort are very welcome.

Outlook

In the last 40 years, structural biology has become highly automated, and methods have advanced to the point that it is now feasible to solve a new structure from start to finish in a matter of months with little specialist knowledge. The extremely rapid and timely solution of these structures is a remarkable achievement during this crisis and, despite some shortcomings, these structures have enabled downstream work on therapeutics to rapidly progress. The downside is that errors at all points during a structure determination are not only common, but can also remain undetected, and if they are detected, this is usually seen as individual failure. However, no individual researcher is fully conversant in all the details of structure determination, protein and nucleic acid structure, chemical properties of interacting groups, catalytic mechanisms, and viral life cycle. The result is that the first draft of a molecular model often contains errors like the ones pointed out above. While any molecular model could benefit from an examination by multiple experts, during this time it is important to bring such inspection to Coronavirus-related structures as quickly as possible.

We believe that, as a community, we need to change how we all see, address and document errors in structures to achieve the best possible structures from our experiments. We are scientists: *In the end, truth should always win.*

Acknowledgements

This work was supported by the German Federal Ministry of Education and Research [grant no. 05K19WWA], Deutsche Forschungsgemeinschaft [project TH2135/2-1], the Wellcome Trust [grants 208398/Z/17/Z and 209407/Z/17/Z], and the US National Institutes of Health [grant R35 GM131883]. It would not have been possible without exchange, discussions and support from the computational and experimental structural biology community. We would particularly like to thank Lu Zhang, John Chodera, Stefano Forli, Thomas Hermanns, Clemens Vonrhein and Arwen Pearson. We are also grateful to Holger Theymann, Nicole Dörfel and Thomas Splettstößer for visualization of our work with their illustration and designs and those who supported us in the last six months: Elisa Bandello, Pairoh Seeliger & Florian Platzmann.

Literature

1. E. N. Baker, Visualizing an unseen enemy; mobilizing structural biology to counter COVID-19. *Acta Cryst D*. **76**, 311–312 (2020).

2. S. K. Burley, H. M. Berman, C. Christie, J. M. Duarte, Z. Feng, J. Westbrook, J. Young, C. Zardecki, RCSB Protein Data Bank: Sustaining a living digital data resource that enables breakthroughs in scientific research and biomedical education. *Protein Science : A Publication of the Protein Society*. **27**, 316 (2018).

3. H. Berman, K. Henrick, H. Nakamura, Announcing the worldwide Protein Data Bank. *NSMB*. **10**, 980 (2003).

4. M. Shirts, V. S. Pande, Screen Savers of the World Unite! *Science*. **290**, 1903–1904 (2000).

5. M. I. Zimmerman, J. R. Porter, M. D. Ward, S. Singh, N. Vithani, A. Meller, U. L. Mallimadugula, C. E. Kuhn, J. H. Borowsky, R. P. Wiewiora, M. F. D. Hurley, A. M. Harbison, C. A. Fogarty, J. E. Coffland, E. Fadda, V. A. Voelz, J. D. Chodera, G. R. Bowman, *bioRxiv*, in press, doi:10.1101/2020.06.27.175430.

6. OpenPandemics - COVID-19 | Research | World Community Grid, (available at https://www.worldcommunitygrid.org/research/opn1/overview.do).

7. JEDI COVID-19 Grand Challenge, (available at https://www.covid19.jedi.group).

8. P. H. Zwart, R. W. Grosse-Kunstleve, P. D. Adams, Xtriage and Fest: automatic assessment of X-ray data and substructure structure factor estimation, 9.

9. A. Thorn, J. Parkhurst, P. Emsley, R. A. Nicholls, M. Vollmar, G. Evans, G. N. Murshudov, AUSPEX: a graphical tool for X-ray diffraction data analysis. *Acta Cryst D*. **73**, 729–737 (2017).

10. J. Rp, L. F, M. Gn, P. A, The PDB_REDO Server for Macromolecular Structure Model Optimization. *IUCrJ*. **1** (2014), , doi:10.1107/S2052252514009324.

11. R. W. W. Hooft, G. Vriend, C. Sander, E. E. Abola, Errors in protein structures. *Nature*. **381**, 272–272 (1996).

12. T. Burnley, C. M. Palmer, M. Winn, Recent developments in the CCP-EM software suite. *Acta Cryst D*. **73**, 469–477 (2017).

13. G. N. Murshudov, P. Skubák, A. A. Lebedev, N. S. Pannu, R. A. Steiner, R. A. Nicholls, M. D. Winn, F. Long, A. A. Vagin, REFMAC5 for the refinement of macromolecular crystal structures. *Acta Cryst D*. **67**, 355–367 (2011).

14. J. Ap, M. S, B. T, W. C, C. Dk, W. M, T. M, Refinement of atomic models in high resolution EM reconstructions using Flex-EM and local assessment. *Methods (San Diego, Calif.)*. **100** (2016), , doi:10.1016/j.ymeth.2016.03.007.

15. P. Mostosi, H. Schindelin, P. Kollmannsberger, A. Thorn, Haruspex: A Neural Network for the Automatic Identification of Oligonucleotides and Protein Secondary Structure in Cryo-Electron Microscopy Maps. *Angewandte Chemie International Edition* (2020), doi:10.1002/anie.202000421.

16. V. B. Chen, W. B. Arendall, J. J. Headd, D. A. Keedy, R. M. Immormino, G. J. Kapral, L. W. Murray, J. S. Richardson, D. C. Richardson, MolProbity: all-atom structure validation for macromolecular crystallography. *Acta Cryst D*. **66**, 12–21 (2010).

17. C. J. Williams, J. J. Headd, N. W. Moriarty, M. G. Prisant, L. L. Videau, L. N. Deis, V. Verma, D. A. Keedy, B. J. Hintze, V. B. Chen, S. Jain, S. M. Lewis, W. B. Arendall, III, J. Snoeyink, P. D. Adams, S. C. Lovell, J. S. Richardson, D. C. Richardson, MolProbity: More and better reference data for improved all-atom structure validation. *Protein Science : A Publication of the Protein Society*. **27**, 293 (2018).

18. M. G. Prisant, C. J. Williams, V. B. Chen, J. S. Richardson, D. C. Richardson, New tools in MolProbity validation: CaBLAM for CryoEM backbone, UnDowser to rethink “waters,” and NGL Viewer to recapture online 3D graphics. *Protein Science*. **29**, 315–329 (2020).

19. L. J, K. Y, H. R, Nsp3 of Coronaviruses: Structures and Functions of a Large Multi-Domain Protein. *Antiviral research*. **149** (2018), , doi:10.1016/j.antiviral.2017.11.001.

20. B. H. Harcourt, D. Jukneliene, A. Kanjanahaluethai, J. Bechill, K. M. Severson, C. M. Smith, P. A. Rota, S. C. Baker, Identification of Severe Acute Respiratory Syndrome Coronavirus Replicase Products and Characterization of Papain-Like Protease Activity. *Journal of Virology*. **78**, 13600 (2004).

21. M. Grabowski, K. M. Langner, M. Cymborowski, P. J. Porebski, P. Sroka, H. Zheng, D. R. Cooper, M. D. Zimmerman, M.-A. Elsliger, S. K. Burley, W. Minor, A public database of macromolecular diffraction experiments. *Acta Cryst D*. **72**, 1181–1193 (2016).

22. C. S. Bury, J. C. Brooks-Bartlett, S. P. Walsh, E. F. Garman, Estimate your dose: RADDOSE-3D. *Protein Science*. **27**, 217–228 (2018).

23. W. Kabsch, XDS. *Acta Cryst D*. **66**, 125–132 (2010).

24. I. J. Tickle, C. Flensburg, P. Keller, W. Paciorek, A. Sharff, C. Vonrhein, G. Bricogne, *STARANISO* (Global Phasing Ltd., Cambridge, 2018; http://staraniso.globalphasing.org/cgi-bin/staraniso.cgi).

25. N. Barretto, D. Jukneliene, K. Ratia, Z. Chen, A. D. Mesecar, S. C. Baker, The Papain-Like Protease of Severe Acute Respiratory Syndrome Coronavirus Has Deubiquitinating Activity. *Journal of Virology*. **79**, 15189 (2005).

26. E. C. Smith, M. R. Denison, Implications of altered replication fidelity on the evolution and pathogenesis of coronaviruses. *Current Opinion in Virology*. **2**, 519 (2012).

27. R. N. Kirchdoerfer, A. B. Ward, Structure of the SARS-CoV nsp12 polymerase bound to nsp7 and nsp8 co-factors. *Nat Commun*. **10**, 1–9 (2019).

28. T. I. Croll, ISOLDE: a physically realistic environment for model building into low-resolution electron-density maps. *Acta Cryst D*. **74**, 519–530 (2018).

29. H. S. Hillen, G. Kokic, L. Farnung, C. Dienemann, D. Tegunov, P. Cramer, Structure of replicating SARS-CoV-2 polymerase. *Nature*, 1–6 (2020).

30. 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, H. E. Xu, Structural basis for inhibition of the RNA-dependent RNA polymerase from SARS-CoV-2 by remdesivir. *Science*. **368**, 1499–1504 (2020).

31. L. Zhang, D. Zhang, C. Yuan, X. Wang, Y. Li, X. Jia, X. Gao, H.-L. Yen, P. P.-H. Cheung, X. Huang, *bioRxiv*, in press, doi:10.1101/2020.04.27.063859.

32. T. A. Steitz, A mechanism for all polymerases. *Nature*. **391**, 231–232 (1998).

33. Y. W. Yin, T. A. Steitz, The Structural Mechanism of Translocation and Helicase Activity in T7 RNA Polymerase. *Cell*. **116**, 393–404 (2004).

34. L. Zhang, R. Zhou, Structural Basis of the Potential Binding Mechanism of Remdesivir to SARS-CoV-2 RNA-Dependent RNA Polymerase. *The Journal of Physical Chemistry B* (2020), doi:10.1021/acs.jpcb.0c04198.

35. K. Barakat, M. Ahmed, Y. Tabana, M. Ha, *bioRxiv*, in press, doi:10.1101/2020.06.02.130849.

36. A. Shannon, N. T.-T. Le, B. Selisko, C. Eydoux, K. Alvarez, J.-C. Guillemot, E. Decroly, O. Peersen, F. Ferron, B. Canard, Remdesivir and SARS-CoV-2: Structural requirements at both nsp12 RdRp and nsp14 Exonuclease active-sites. *Antiviral Research*. **178**, 104793 (2020).

37. A. Krylov, T. L. Windus, T. Barnes, E. Marin-Rimoldi, J. A. Nash, B. Pritchard, D. G. A. Smith, D. Altarawy, P. Saxe, C. Clementi, T. D. Crawford, R. J. Harrison, S. Jha, V. S. Pande, T. Head-Gordon, Perspective: Computational chemistry software and its advancement as illustrated through three grand challenge cases for molecular science. *The Journal of Chemical Physics*. **149**, 180901 (2018).

38. COVID-19 Molecular Structure and Therapeutics Hub, (available at https://covid.molssi.org/).