Making the invisible enemy visible

Structural biology plays a crucial role in the fight against COVID-19, permitting us to 'see' and understand SARS-CoV-2. However, the macromolecular structures of SARS-CoV-2 proteins that were solved with great speed and urgency can contain errors that may hinder drug design. The Coronavirus Structural Task Force has been working behind the scenes to evaluate and improve these structures, making the results freely available at https://insidecorona.net/.

Tristan I. Croll, Kay Diederichs, Florens Fischer, Cameron D. Fyfe, Yunyun Gao, Sam Horrell, Agnel Praveen Joseph, Luise Kandler, Oliver Kippes, Ferdinand Kirsten, Konstantin Müller, Kristopher Nolte, Alexander M. Payne, Matthew Reeves, Jane S. Richardson, Gianluca Santoni, Sabrina Stäb, Dale E. Tronrud, Lea C. von Soosten, Christopher J. Williams and Andrea Thorn

When the COVID-19 pandemic hit in early 2020, the structural biology community quickly swung into action to determine the atomic structures of the 28 viral proteins encoded by SARS-CoV-2 (ref. [1]). A total of 1,146 structures covering 18 SARS-CoV-1 and SARS-CoV-2 proteins have been released over the course of just 12 months. They are freely and publicly available in the World Wide Protein Data Bank (wwPDB), which celebrates its 50th anniversary this month. These models serve as the basis for structure-based drug design and vaccine development. They are also essential for understanding how the virus hijacks human cells and causes disease. However, errors occur in even the most carefully determined structures and are probably more common in structures solved quickly and under immense pressure. Even small errors can have severe consequences for structure-based drug discovery, structural bioinformatics and computational chemistry because they can be misinterpreted as biologically and pharmaceutically relevant.

While the wwPDB is an invaluable tool, serving as structural biology's archive of record, it is also largely static. Released structures can only be updated by the original depositors, and there is often little motivation to make corrections once associated papers are published. 99% of PDB structure downloads are not conducted by experimental structural biologists but by scientists who use the structural data[2] and who may lack the training to identify and correct erroneous sites in the molecular model.

In this global crisis, it is vital to ensure that the available structural data are the best they can be, which requires us to push our methods to the limit. The Coronavirus Structural Task Force, a diverse international team of structural biologists involved in

methods development, responded to this challenge by rapidly categorizing, evaluating and reviewing all experimental protein structures of SARS-CoV-1 and SARS-CoV-2, which comprise the subgenus *Sarbecovirus*. We do a weekly automatic post-analysis as well as a manual reprocessing and remodeling of representative structures from each of the 18 structurally characterized *Sarbecovirus* proteins. Every Wednesday, when new PDB structures are released, our automated pipeline identifies new coronavirus structures and assesses the quality of the models and experimental data. This assessment, along with the original structures, is immediately made available in our online repository at https://insidecorona. net/. There we also supply a summary, an

SQL database of key statistics and quality indicators, and individual results. After our validation effort began, researchers involved in in silico drug screening from Folding@Home[3], OpenPandemics[4] and the EU Joint European Disruptive Initiative (JEDI) expressed great interest. These groups aim to simulate the conformational flexibility of coronavirus proteins and their interactions with each other and with host cell proteins and to design small-molecule inhibitors against key SARS-CoV-2 targets via high-throughput computational modelling, a task that is exquisitely sensitive to the quality of the input model.

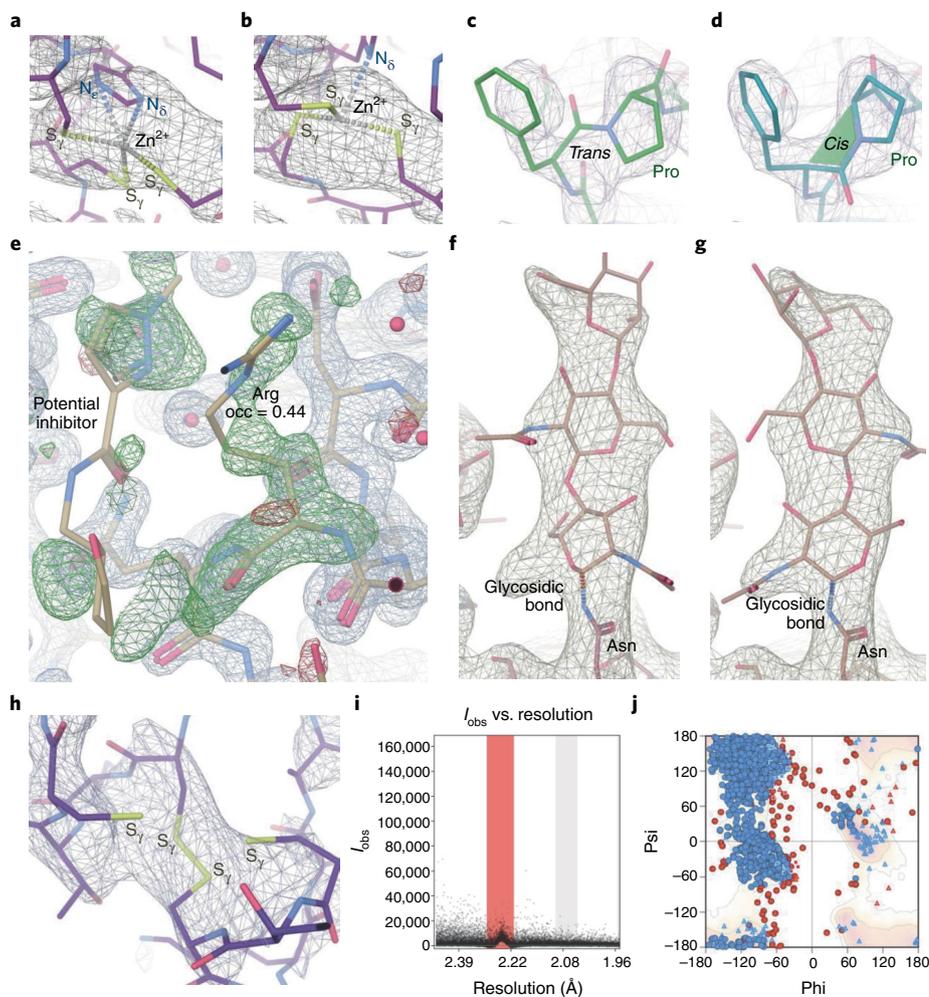In addition to structure evaluation and improvement, https://insidecorona.net/

Table 1 | Examples of quality indicators pointing to potential problems in PDB entries, calculated using our automatic evaluation pipeline

Key indicators in evaluation	Number of depositions (Percentage of total)
X-ray crystallography (999 depositions)	
Completeness <80%	14 (1.4%)
R_{free} >35%	2 (0.2%)
Potential twinning	52 (5.2%)
Contaminated by ice diffraction	93 (9.3%)
Incorrect mask	86 (8.6%)
Single-particle cryo-EM (360 depositions)	
Average model–map FSC <0.4	46 (13%)
MI score <0.4	56 (16%)
SMOC >10%	64 (18%)
Other indicators (1,392 depositions, including 31 NMR and 2 neutron diffraction)	
CaBLAM outlier conformations >2.0%	318 (23%)
CaBLAM severe C_{α} outliers >1.0%	122 (8.8%)
Sequence mismatch	23 (1.7%)

Potential twinning was identified by L-Test[7], and ice diffraction and incorrect mask were identified by visual inspection of AUSPEX plots[8]. The chosen cutoffs for FSC and MI score[16] indicate poor overall agreement between the map and model. A SMOC[15] score of >10% indicates that more than 10% of the residues of a structure fit poorly with the map and could potentially be improved.

Fig. 1 | Examples of common errors and improvements. All pictures except **i** are screenshots from the Coot v0.9.9 prerelease. Residual density and reconstruction maps are in blue-gray, difference electron density in red and green. **a**, SARS-CoV-1 Nsp14–Nsp10 (PDB 5C8T) histidine zinc-coordination site (B603), with residual density contour level 0.445, root mean square deviation (r.m.s.d.) 0.150. **b**, Histidine from **a** has been swapped in ISOLDE[25], leading to tetrahedral coordination of Zn^{2+}, then refinement was performed using PDB-REDO[11] with manual addition of links. **c**, Proline A505 is modelled as *trans* in the RdRp complex (PDB 7BV2, left), but the density indicates a *cis* main chain conformation, shown in **d**. **d**, The deposited PDB entry was updated after we contacted the original authors. **e**, High difference electron density at residue A165 in the SARS-CoV-2 main protease (PDB 5RFA) due to an occupancy of only 0.44 rather than 1.00 near the potential inhibitor (left). Residual map contour level 0.54, r.m.s.d. 0.319; difference density at contour level 0.35, r.m.s.d. 0.114. **f**, SARS-CoV-2 spike receptor-binding domain complexed with human ACE2 (PDB 6VW1). This *N*-linked glycan is flipped approximately 180° around the *N*-glycosidic bond. After we contacted the original authors, this entry was revised (shown in **g**). **g**, Correction improves the density fit of the sugar chain. Residual map at contour level 0.311, r.m.s.d. 0.265. **h**, Disulfide bond A226–A189 in papain-like protease (PDB 6W9C), with electron density at contour level 0.214, r.m.s.d. 0.136; the other two cysteine residues remain uncoordinated. While the density map does not indicate a zinc, it is a zinc finger domain; the other NCS copies include a coordinated zinc at this position. **i**, AUSPEX[8] plot of SARS-CoV main protease (PDB 2HOB); ice rings are reflected by a bias in the intensity distribution (red). **j**, Ramachandran plot or torsion angles in the peptide backbone for the SARS-CoV Nsp10–Nsp14 dynamic complex (PDB 5NFY). In principle, there should only be a few outliers (red), as most peptide bonds adhere to typical angular distributions. Picture: CSTF/insidecorona.net.

supplies literature reviews that discuss the structural aspects of the viral infection cycle and host interaction partners and

provides advice on selecting the best starting models for in silico projects. Furthermore, we have added SARS-CoV-2 proteins to

Proteopedia[5] and http://molssi.org/ and we have included a 3D-Bionotes[6] deep link in our database. Finally, we have tried to make SARS-CoV-2-related research accessible to the general public with blog posts aimed at non-scientists. We also live streamed data processing on Twitch and provided an accurate 3D printed model of SARS-CoV-2 that is based on deposited structures, along with the files and instructions necessary to print these models.

Automatic evaluation
All macromolecular *Sarbecovirus* structures in the wwPDB are downloaded into our repository and assessed automatically within 24 hours of their release. We combine new validation tools with previously developed methods, many of which were adapted for our purposes.

Crystallographic data and structure solutions. 73% of the 1,392 reported *Sarbecovirus* structures were derived by X-ray crystallography. We evaluate these datasets for pathologies such as twinning, multiple lattice diffraction, ice crystal contamination, incompleteness and radiation damage using phenix.xtriage[7] and AUSPEX[8]. Although these issues cannot be resolved after data collection, taking them into account during data processing and structure solution can yield better models. It can be difficult to identify these problems using deposited structure factors, since information is lost during the processing of raw diffraction data. Raw data allow a more complete analysis of the experiment and reprocessing but can be difficult to obtain, as they are neither deposited in the wwPDB nor required for publication. We therefore invite authors to send us their raw experimental data and offer to deposit them in public repositories, such as SBGrid[9] or https://proteindiffraction.org/[10]. All data sets we have analyzed to date have an acceptable signal-to-noise ratio; we have also evaluated other statistical quality indicators, examples of which are summarized in Table 1.

A general indication of how well the atomic model fits the measurement data is given by the R values. While only two structures in our database present alarmingly high R_{free} values, that is, above 35%, this does not necessarily mean there are no modelling problems. Large R_{free} drops indicate major issues with PDB entries, especially for older SARS-CoV-1 structures. PDB-REDO[11] re-refinements generally improved R_{free}. Nevertheless, the resulting models should not be viewed as "more correct" purely on the basis of a lower R value, particularly at lower resolution, where the relationship between R values and
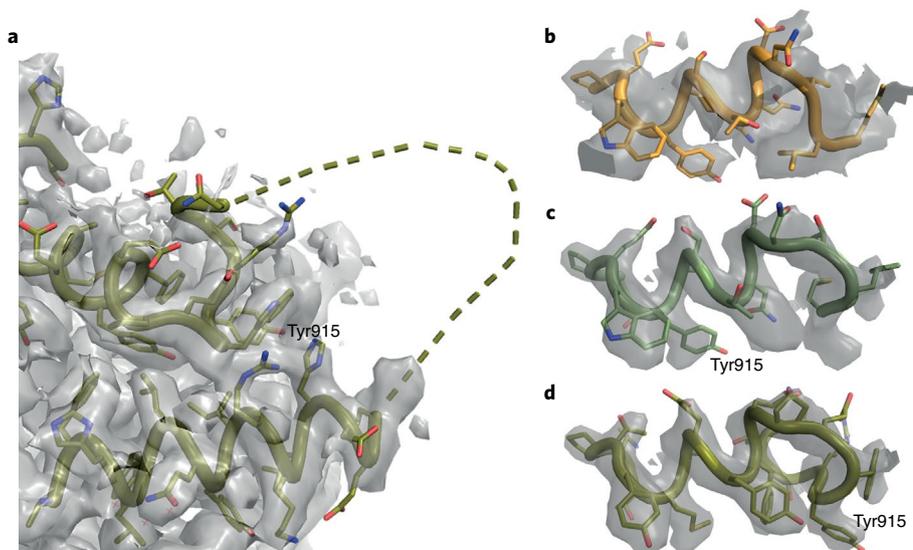
Fig. 2 | Register shift in the C terminus of RNA polymerase. a, Overview with missing loop shown as a dashed line (PDB 7BV2); map at 2.4σ. Right, details of the C-terminal helix at 5σ. **b**, Lower resolution map and model (PDB 6NUS). Judging the side chain fit is difficult. **c**, Higher resolution map and model (PDB 7BV2) as deposited; the side chain fit is suboptimal due to the register error. **d**, Amended model for PDB 7BV2; the side chains now fit the density. The register shift is indicated by the labelled Tyr915. Picture: CSTF/insidecorona.net.

model quality degrades[12]. Critical manual inspection of the model remains necessary.

Structures from single-particle cryo-EM.
Cryo-EM structures make up 24% of reported SARS-CoV-1 and SARS-CoV-2 structures. Raw data are not available from the wwPDB, but deposition into EMPIAR[13] is increasingly common. The reconstructed 3D map deposited in the EMDB[14] allows calculation of the fit between model and map using Fourier shell correlation (FSC) to assess the agreement between features at different resolutions. FSCs, real-space cross-correlation coefficient (CCC), mutual information (MI) and segment Manders' overlap coefficient (SMOC)[15] were calculated with the CCP-EM[16] model validation task (Table 1). While MI and CCC are single-value scores that indicate how well the model and map agree overall, the SMOC score evaluates the fit of each modelled residue individually and can highlight specific regions where the model and map disagree. We use Haruspex[17], a neural network trained to recognize secondary structure elements and RNA and DNA in cryo-EM maps, to provide visual guidance for manual structure evaluation.

Evaluation of the structural models using prior knowledge. MolProbity[18] is used to evaluate the model quality and to check covalent geometry and conformational parameters of protein and RNA and steric

clashes. Some of these traditional quality indicators are used as additional restraints during refinement, which reduces their usefulness as quality metrics. The newer MolProbity CaBLAM score[6] is designed to find local errors and is particularly useful at 3–4 Å resolution. Current refinement packages do not specifically aim at improving this score, arguably making it a more reliable quality indicator. In addition, checking the amino acid sequence of each model against that in the deposited PDB file highlighted mismatches in 23 cases. During the COVID-19 crisis, the MolProbity web service has been pushed to its limit as drug developers screen the same SARS-CoV-2 structures many times. We developed a custom MolProbity pipeline that makes the validation results for these structures available online, thereby decreasing the web service's workload.

Manual evaluation
Although the structural biology community has achieved a high level of automation in data collection, data processing and structure solution in recent years, the process of structure determination still requires interpretation by researchers. This especially applies to low-quality maps with poor fit between experimental data and structural models. Visual residue-by-residue inspection by an experienced structural biologist remains the best way to judge quality. We therefore select representative

structures of each SARS-CoV-2 protein, as well as those of particular interest for drug development, for manual evaluation. Certain problems are surprisingly common, such as peptide bond flips (Fig. 1c,d), rotamer errors, occupancy problems (Fig. 1e) and misidentification of small molecules or ions, for example, water as magnesium and chloride as zinc. Of note, zinc plays an important role in many SARS-CoV-2 proteins. We found many zinc coordination sites to be mismodelled, with the zinc ion missing or pushed out of the density and/or erroneous disulfide bonds between the coordinating cysteine residues (Fig. 1a,b,h). In addition, many coronavirus proteins are glycosylated at surface asparagine residues, but glycan sugars were often flipped from their correct orientation around the N-glycosidic bond (Fig. 1f,g). This can be avoided by using tools such as Privateer[19] and the automated carbohydrate building tool in Coot[20]. It is important to note that deviation from expected behavior is not always an error and can also be a functionally relevant feature, for example, the strained geometries often found at catalytic sites. However, such deviations must be strongly supported by the experimental data. Of the structures we checked manually, we were able to substantially improve 31 in terms of model quality, data quality, or both. Below we give two examples to illustrate the importance of carefully inspecting the experimental data and resulting models.

Papain-like protease. SARS-CoV-2 nonstructural protein 3 (Nsp3) contains a papain-like protease domain that is essential for infection because it cleaves the viral polypeptide. The first structure of the SARS-CoV-2 papain-like protease (PDB 6W9C) was released 1 April 2020, only three months after the viral genome was reported (GenBank MN908947.2)[21]. The structure was immediately used in drug design efforts. The overall completeness of the measured data, however, was only 57%. Examination of the raw data, available from https://proteindiffraction.org/[10], revealed strong radiation damage, exacerbated by a poor data collection strategy. This could not be deduced from the PDB deposition, underlining the importance of making raw data available.

The crystal has 3-fold non-crystallographic symmetry (NCS), with each papain-like protease domain monomer containing a functionally important Zn^{2+} ion bound by four cysteine residues with similar C_β–S_γ–Zn angles and Zn–S_γ bond lengths. Because of radiation damage, the Zn–S sites have poor density. In one NCS copy, the site

has been modelled as a disulfide bond and two free cysteine residues (Fig. 1h), while the other two NCS copies coordinate the zinc atom with strongly varying C_β–S_γ–Zn angles and Zn–S bond lengths. We reprocessed the images using XDS[22], a software for the processing of single-crystal X-ray diffraction images. The STARANISO server was used to determine and apply an anisotropic limit for the diffraction data. This careful manual intervention improved the overall quality of the data and increased the resolution from 2.7 to 2.6 Å, but the revised overall ellipsoidal completeness was only 44.5%. Adding zinc atoms to all sites, restraining the bond lengths and angles to the expected values and using NCS restraints and an overall higher weighting for ideal geometry, together with remodeling the side chains and water molecules, improved the electron density maps and lowered the R values by 4%. This exemplifies the interconnection between data collection, data processing and model building: even if the data collection strategy is not ideal, taking the resulting problems into account during data processing and refinement can drastically improve the final model.

A structure of the C111S mutant of the papain-like protease domain (PDB 6WRH) was released one month later. In this structure, the zinc sites were clearly resolved in all subunits. In the meantime, however, PDB 6W9C had been widely used in in silico drug design. 20% of the over 140 research teams in the JEDI COVID19 GrandChallenge, a competition to find potential COVID-19 drugs in silico, have used this model. The availability of a better structure one month earlier would have increased their chances of success and saved computing and person hours.

RNA polymerase complex. SARS-CoV-2 replicates its single-stranded RNA genome using a macromolecular complex of RNA-dependent RNA polymerase (Nsp12; RdRp), Nsp7 and Nsp8. Earlier cryo-EM structures of the SARS-CoV-1 homologues (PDB 6NUR, PDB 6NUS) include a disordered unmodelled loop followed by a visible but short and irregular helix and a flexible C terminus. Density for this helix was poorly resolved, but the model had valid geometry. Our analysis of one of the first structures of the equivalent SARS-CoV-2 complex (PDB 7BTF) revealed that the sequence in this C-terminal region (part of the RNA-binding groove) was misaligned by nine residues (Fig. 2). This error was present in all related SARS-CoV-1 and SARS-CoV-2 structures, probably because new structure determination typically starts from an earlier model when one is available.

A structure of the RdRp complex bound to the nucleotide analogue remdesivir (PDB 7BV2 (ref. [23])) was released soon after and provided the basis for rational design of related drug candidates[24]. This structure also featured the nine-residue sequence misalignment. We rebuilt the structure using ISOLDE[25], CaBLAM[6] and visual inspection, correcting some flipped or *cis* versus *trans* peptides (Fig. 1c,d) and three RNA conformers near remdesivir, including a backward adenosine base. We were also able to add several residues and waters with good density and geometry. Remdesivir is covalently attached to the RNA, but it is only present in an estimated $\leq$50% of the measured molecules[12]. This means that the active site is a mixture of at least two different states, so unsurprisingly, the modeled Mg^{2+} ions and pyrophosphate are poorly supported by the experimental density and local contacts. This is of concern for subsequent in silico docking and drug design, which often take all atoms in the deposited structure as a fixed framework to build into. The remodelled structures of the complex may offer a more solid basis for drug design, even if the ~50% occupancy of the active site was not widely discussed[12]. It is notable that despite the large register error and various smaller issues, by traditional "summary" metrics the model appeared extremely good, with no Ramachandran nor rotamer outliers and a clash score of 2, highlighting that direct visual inspection must remain a key step in any modelling process.

Although the problems discussed above were present in the originally deposited structures, nearly all are now corrected. This was achieved at least in part because we made corrected models available on our website and contacted the original authors of these structures with detailed descriptions, supporting them to deposit revised versions to the wwPDB at their discretion.

Conclusion
In the past 50 years, structural biology has achieved a high level of automation, and methods have advanced greatly. It is now feasible to solve a new structure from start to finish in a matter of weeks, with little specialist knowledge. This is exemplified by the rapid solution of SARS-CoV-2 structures during the pandemic, which is a remarkable achievement. These structures have enabled rapid progress in the development of therapeutics and vaccines. However, errors at all stages of structure determination are not only common but often remain undetected. Unfortunately, no individual researcher can be fully conversant in all of the details of structure determination,

the chemical properties of interacting groups, catalytic mechanisms and the viral infection cycle. While any molecular model could benefit from examination by multiple experts, it is particularly important to rapidly carry out such inspection of coronavirus-related structures in the context of the current pandemic.

Structural models are an interpretation of the measured data, and deposited structures should be seen as an initial interpretation that can provide considerable biological insight but may leave room for improvement. The availability of raw data would allow a more complete assessment of the structure solution. It would also offer the opportunity to reanalyze the data and to propose updates to the original authors or to deposit derivative models in the wwPDB. We believe that, as a community, we need to change how we 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. ❏

Tristan I. Croll[1], Kay Diederichs[2], Florens Fischer[3,5], Cameron D. Fyfe[4], Yunyun Gao[3,5], Sam Horrell[6], Agnel Praveen Joseph[7], Luise Kandler[3,5], Oliver Kippes[3,5], Ferdinand Kirsten[3,5], Konstantin Müller[5], Kristopher Nolte[3,5], Alexander M. Payne[8], Matthew Reeves[5], Jane S. Richardson[9], Gianluca Santoni[10], Sabrina Stäb[3,5], Dale E. Tronrud[11], Lea C. von Soosten[3,5], Christopher J. Williams[9] and Andrea Thorn[3,5]✉

[1]CIMR, University of Cambridge, Cambridge, UK. [2]Universität Konstanz, Konstanz, Germany. [3]Institut für Nanostruktur und Festkörperphysik, Universität Hamburg, Hamburg, Germany. [4]Micalis Institute, INRAE, Jouy-en-Josas, France. [5]Rudolf -Virchow-Zentrum, Julius-Maximilians-Universität Würzburg, Würzburg, Germany. [6]Diamond Light Source, Didcot, UK. [7]Science and Technology Facilities Council, Swindon, UK. [8]Memorial Sloan Kettering Cancer Center, New York, NY, USA. [9]Duke University, Durham, NC, USA. [10]European Synchrotron Radiation Facility, Grenoble, France. [11]Oregon State University, Corvallis, OR, USA.
✉e-mail: andrea.thorn@uni-hamburg.de

References
1. Baker, E. N. *Acta Cryst. D Struct. Biol.* **76**, 311–312 (2020).
2. Burley, S. K. et al. *Protein Sci.* **27**, 316–330 (2018).
3. Zimmerman, M. I. et al. Preprint at *bioRxiv* https://doi.org/10.1101/2020.06.27.175430 (2020).
4. World Community Grid. Research. *OpenPandemics—COVID-19* https://www.worldcommunitygrid.org/research/opn1/overview.do (2021).
5. Prilusky, J. et al. *J. Struct. Biol.* **175**, 244–252 (2011).
6. Prisant, M. G., Williams, C. J., Chen, V. B., Richardson, J. S. & Richardson, D. C. *Protein Sci.* **29**, 315–329 (2020).

7. Zwart, P. H., Grosse-Kunstleve, R. W. & Adams, P. D. *CCP4 Newsletter* http://legacy.ccp4.ac.uk/newsletters/newsletter43.pdf (2005).
8. Thorn, A. et al. *Acta Cryst. D Struct. Biol.* **73**, 729–737 (2017).
9. Morin, A. et al. *Elife* **2**, e01456 (2013).
10. Grabowski, M. et al. *Acta Cryst. D Struct. Biol.* **72**, 1181–1193 (2016).
11. Joosten, R. P., Long, F., Murshudov, G. N. & Perrakis, A. *IUCrJ* **1**, 213–220 (2014).
12. Croll, T. I., Williams, C. J., Chen, V. B., Richardson, D. C. & Richardson, J. S. *Biophys. J.* **120**, 1085–1096 (2021).
13. Iudin, A., Korir, P. K., Salavert-Torres, J., Kleywegt, G. J. & Patwardhan, A. *Nat. Methods* **13**, 387–388 (2016).
14. Lawson, C. L. et al. *Nucleic Acids Res.* **39**, D456–D464 (2011).
15. Joseph, A. P. et al. *Methods* **100**, 42–49 (2016).
16. Burnley, T., Palmer, C. M. & Winn, M. *Acta Cryst. D Struct. Biol.* **73**, 469–477 (2017).
17. Mostosi, P., Schindelin, H., Kollmannsberger, P. & Thorn, A. *Angew. Chem. Int. Ed. Engl.* **59**, 14788–14795 (2020).
18. Williams, C. J. et al. *Protein Sci.* **27**, 293–315 (2018).
19. Agirre, J. et al. *Nat. Struct. Mol. Biol.* **22**, 833–834 (2015).
20. Emsley, P. & Crispin, M. *Acta Cryst. D Struct. Biol.* **74**, 256–263 (2018).
21. Wu, F. et al. *Nature* **579**, 265–269 (2020).
22. Kabsch, W. *Acta Cryst. D Struct. Biol.* **66**, 125–132 (2010).
23. Yin, W. et al. *Science* **368**, 1499–1504 (2020).
24. Zhang, L. et al. *Phys. Chem. Chem. Phys.* **23**, 5852–5863 (2021).
25. Croll, T. I. *Acta Cryst. D Struct. Biol.* **74**, 519–530 (2018).

Author contributions
K.N. did sequence alignments and PDB downloads; the automatic evaluation pipeline and SQL database were designed and set up by G.S., Y.G., K.N., A.T., O.K. and K.M.; weekly updates are run by Y.G., G.S. and K.N. The pipeline contains software by A.T., A.P.J., J.S.R., C.J.W., Y.G. and K.N., as well as many external collaborators, such as R. Joosten. The survey of each new structure was coordinated by Y.G. and done by F.F., C.D.F., S.H., L.K., O.K., F.K., K.N., G.S., S.S., L.C.v.S. & A.T. Zinc coordination sites were analyzed by M.R. AUSPEX plots were interpreted by S.S. Manual reprocessing and rebuilding of structures was done by T.I.C., D.E.T., S.H., A.P.J., C.D.F., A.T. and K.D. Reviews and illustrations of *Sarbecovirus* structures on the homepage were provided by C.D.F., S.H., L.K., O.K., F.K., K.N., A.M.P., M.R., G.S., S.S., D.E.T., L.C.v.S., A.T., T. Splettstößer (SciStyle.com) and N. Dörfel. 3D models were designed by D.E.T. with help from T. Splettstößer and were made by K.N., M.R., L.K., S.S. and D.E.T. A.T. conceived and supervised the project. All authors wrote the manuscript together.

Competing interests
The authors declare no competing interests.