A regional outbreak of pneumonia in Wuhan, Hubei Province of China in late 2019 was associated with a novel coronavirus. Rapid release of genomic data for the isolated virus enabled the construction of first-generation homology models of the new CoV 3CL<sup>pro</sup> cysteine protease. Whilst the overall viral genome was most closely associated with bat coronaviruses, the main protease is most closely related (96% identity) to SARS CoV protease.
Homology models of Wuhan Coronavirus 3CL$^{\text{pro}}$ Protease

Martin J. Stoermer*
Division of Chemistry and Structural Biology, Institute for Molecular Bioscience, The University of Queensland, St. Lucia, 4072, Queensland, Australia.

*Corresponding author
Dr. Martin J. Stoermer
Division of Chemistry and Structural Biology, Institute for Molecular Bioscience The University of Queensland
St. Lucia, 4072, Queensland, Australia
E-mail: martin.stoermer@uq.edu.au
Orcid: https://orcid.org/0000-0003-3445-2104
Twitter: @MartinStoermer

Abstract.

A regional outbreak of pneumonia in Wuhan, Hubei Province of China in late 2019 was associated with a novel coronavirus. Rapid release of genomic data for the isolated virus enabled the construction of first-generation homology models of the new CoV 3CL$^{\text{pro}}$ cysteine protease. Whilst the overall viral genome was most closely associated with bat coronaviruses, the main protease is most closely related (96% identity) to SARS CoV protease.

Keywords: Wuhan coronavirus, 3CL$^{\text{pro}}$, protease, homology modelling
Foreword and Acknowledgements

The author is indebted to Professor Yong-Zhen Zhang, Fudan University, Shanghai, the Shanghai Public Health Clinical Center & School of Public Health, the Central Hospital of Wuhan, Huazhong University of Science and Technology, the Wuhan Center for Disease Control and Prevention, the National Institute for Communicable Disease Control and Prevention, the Chinese Center for Disease Control and Prevention, Edward Holmes, The University of Sydney, Sydney, Australia for the rapid provision of the Wuhan coronavirus genome sequence WH-Human_1 and subsequent GenBank submissions to the international research community and the public.

Introduction.

In December 2019, the World Health Organisation (WHO) was officially informed of a cluster of cases of pneumonia of unknown cause detected in Wuhan City, Hubei Province of China. Subsequently researchers from the Shanghai Public Health Clinical Center & School of Public Health reported that the evidence suggested an origin in a seafood market in Wuhan, and the isolation of a new type of coronavirus (novel coronavirus, nCoV) on 7 January 2020. Laboratory testing ruled out other respiratory pathogens including Severe Acute Respiratory Syndrome coronavirus (SARS-CoV).

On 10th January 2020 a preliminary sequence (WH-Human_1.fasta.gz) of the novel coronavirus was made available on virological.org, and subsequently on 12th January 2020, was released on Genbank (ID MN908947). Preliminary analysis of the Wuhan virus suggests it is most closely related to coronaviruses found in bats.

Proteases are an important class of drug target, most successfully in the cases of HIV protease and Hepatitis C. Recent human outbreaks of related coronaviruses SARS in 2002 and MERS in 2012 have focused the world’s attention on the global risks of such emerging viruses. Coronaviruses are positive strand, enveloped RNA viruses with exceptionally large genomes which usually encode two or three viral proteases. In the case of Severe Acute Respiratory Syndrome coronavirus, the two proteases are a papain-like cysteine protease (PLpro) and a chymotrypsin-like cysteine protease 3CLpro, also known as the Main protease. SARS-CoV 3CLpro is essential for viral replication and is thus recognized as a potential drug target for the treatment of SARS infections. To date there are no clinically used inhibitors of the SARS protease yet they remain in development.
Figure 1. X-ray crystal structure of SARS-CoV 3CL\textsuperscript{\textpro} bound to a covalent inhibitor (PDB code 2a5i). Panel A: Domain I shown as yellow ribbons, domain II cyan ribbons, domain III pale pink ribbons. Catalytic dyad His41, Cys145 shown as yellow sticks, covalent azapeptide inhibitor shown as grey sticks. Panel B: Active site at the interface of domains I, II with P1 Gln residue shown hydrogen bonding to S1 subsite residues Phe140, His163, and Glu166 (blue lines).

The active site of SARS-CoV 3CL\textsuperscript{\textpro} is located in the cleft between domains I and II of the protein, and the two domains each contribute one residue to the catalytic dyad composed of His41 and Cys145 (Figure 1). Substrates and inhibitors typically contain Gln residues at the P1 position and large hydrophobic residues Phe/Leu/Met at P2.\textsuperscript{15} The wealth of medicinal chemistry and biochemical information gained over the last 18 years, coupled with good-high sequence similarity between the members of the coronavirus family makes the homology modelling of new and emerging viruses an inviting prospect.

Results and Discussion.

The release of the viral genome of the novel Wuhan coronavirus prompted us to construct first-generation homology models of the new CoV 3CL\textsuperscript{\textpro}. Initial studies and BLAST analysis indicated that the virus was most closely related to a clade of bat coronaviruses (Figure 2), of which the closest neighbour for which there was protease structural information available in the Protein Data Bank (www.rcsb.org) was bat coronavirus HKU4.\textsuperscript{16,17} EMBOSS Needle\textsuperscript{18} alignment of the Wuhan genome with the sequence extracted from PDB:2YNB identified a putative 306ss viral 3CL\textsuperscript{\textpro} protein with 49\% sequence identity (65\% sequence similarity) (Figure 3).
Figure 2. Phylogenetic Neighbour-joining tree without distance corrections for Wuhan coronavirus and top BLAST hits generated with Clustal Omega.
**Figure 3.** EMBOSS Needle alignment of the extracted protease sequence of Wuhan coronavirus against Bat coronavirus (PDB: 2ynb) and SARS (PDB: 2z9j). Catalytic dyads shown in red bold.

An initial Swissmodel search for possible homology model templates against this search sequence however provided only SARS protease templates with much higher sequence identity and similarity (96, 99% respectively). This was confirmed via a second EMBOSS Needle alignment (Figure 3), and visually by a mapping of differing residues onto the respective crystal structures (Figure 4).

**Figure 4.** Visualisation of sequence differences between bat HKU4, SARS and Wuhan coronaviruses. Panel A: X-ray crystal structure (PDB: 2z9j) of SARS CoV protease shown as green ribbons, with catalytic His41 and Cys145 residues shown as cyan sticks. Residues which differ in the Wuhan sequence are marked in magenta. Panel B: X-ray crystal structure (PDB: 2ynb) of bat coronavirus HKU4 protease shown as green ribbons, with residues which differ in the Wuhan sequence are marked in magenta.
Alignment of the SARS 2z9j and HKU4 2ynb protease crystal structures however showed that the two had a very similar overall fold with an RMSD of 0.66Å across 229 well conserved Cα atoms. Accordingly, due to their higher overall sequence identities, SARS crystal structures were chosen for the future development of Wuhan CoV protease homology models. Coronavirus crystal structures are predominantly homodimers, and as a consequence Swissmodel template searches returned some duplicates arising from both chains A&B being selected. Models were built as homodimers where appropriate however for visualisation and initial refinement purposes the monomers were used (Table 1, Figure 5A). Pairs 1&3, 4&5 were functionally identical so only 6 models were taken through for further refinement. The 5 Swissmodel dimer models 1,2,4,6,7 were refined as both dimers and monomers, using Schrodinger Suite 2019 and monomer model 8 was refined without reconstruction to a dimer. The models were overlaid as their monomers and compared by manual inspection. The models were very structurally similar, especially in the active site region near His41 and Cys145 (Figure 5B).

### Table 1. PDB templates used for Swissmodel homology model building. Templates ranked by internal Swissmodel QSQE score

| Model | Template name | Title                | % Identity | QSQE score | Method (Resolution) | Oligo State | Ligands |
|-------|---------------|----------------------|------------|------------|---------------------|-------------|---------|
| 1     | 2z9j.1.B      | 3C-like proteinase   | 96.1       | 0.91       | X-ray, 2.0Å         | homodimer   | 2x DTZ |
| 2     | 3vb3.1.A      | 3C-like proteinase   | 96.1       | 0.91       | X-ray, 2.2Å         | homodimer   | None   |
| 3     | 2z9j.1.A      | 3C-like proteinase   | 96.1       | 0.91       | X-ray, 2.0Å         | homodimer   | 2x DTZ |
| 4     | 1uk3.1.B      | 3C-like proteinase   | 96.1       | 0.88       | X-ray, 2.4Å         | homodimer   | None   |
| 5     | 1uk3.1.A      | 3C-like proteinase   | 96.1       | 0.88       | X-ray, 2.4Å         | homodimer   | None   |
| 6     | 2a5i.1.A      | 3C-like peptidase    | 96.1       | 0.86       | X-ray, 1.9Å         | homodimer   | 2x AZP |
| 7     | 1uj1.1.B      | 3C-like proteinase   | 96.1       | 0.86       | X-ray, 1.9Å         | homodimer   | None   |
| 8     | 1zl1.1.A      | 3C-like proteinase   | 96.1       | -          | X-ray, 2.8Å         | monomer     | None   |

**Figure 5.** Panel A: Overlay of the 8 Wuhan coronavirus protease monomers produced via Swissmodel. Panel B: Overlay of the 6 energy minimised Wuhan coronavirus protease monomers produced in Schrödinger Suite. Model1 from 2z9j, green ribbons, model2 from 3vb3 cyan, model4 from 1uk3 magenta, model6 from 2a5i yellow, model7 from 1uj1 pale brown, and model8 from 1zl1 grey ribbons.

By a combination of Swissmodel scoring of the initial models and MolProbity analysis of the refined overall model quality, the final two models selected were Model2 from PDB:3vb3 representing the unbound or apo form of the protease, and Model6 from PDB:2a5i which represents a ligand-bound form. The original SARS CoV 3CL\textsuperscript{pro} inhibitor was carried through the modelling protocol and is shown in the active site of the Wuhan 3CL\textsuperscript{pro} model (Figure 5). The apo-like and bound-like models 2,4 are very similar overall except for a significant
difference in the loop 4TSEDM which forms the boundary of the S2 subsite. This results in the bulky Met49 side chain moving out of the S2 pocket allowing the ligand Phe residue to be accommodated.

**Figure 5.** Overlay of ligand-bound and unbound forms of Wuhan coronavirus 3CLpro. Panel A: Apo Model2 from PDB:3vb3 shown as green ribbons, bound Model6 from PDB:2a5i cyan ribbons. Catalytic triad yellow sticks, SARS CoV 3CLpro inhibitor shown as grey sticks. Panel B: Active site of the overlaid bound and unbound forms showing Met49 and mobile loop in S2 pocket.

These models were also analysed using 100ns molecular dynamics simulations (see Supporting Information) to check the systems for overall structural stability (Figure 6). In general the dimeric forms were slightly more stable overall, as the C- and N-termini of the chains form part of the dimer interface and are held more closely than in the monomeric state where they are more mobile.

**Figure 6.** Comparison of mobile regions in molecular dynamics simulations. Panel A: Monomeric model2 as green ribbons except magenta ribbons for moderately mobile regions. Catalytic dyad yellow sticks. Panel B: Dimeric model12 shown as green ribbons Chain A, cyan ribbons Chain B, except magenta ribbons for moderately mobile regions. Catalytic dyad green and cyan sticks.
Finally an analysis of the new Wuhan coronavirus genome revealed multiple sites, analogous to those found or proposed for SARS\textsuperscript{14} that may be cleaved by the 3CLpro after (L/F/M)-Gln residues. These are mostly identical to the SARS sequences (Table 2).

| SARS 3CL\textsuperscript{pro} cleavage site | Site | Potential Wuhan coronavirus sites |
|--------------------------------------------|------|----------------------------------|
| AVLQ ↓ SGFR                               | TM2/3CLpro | AVLQ ↓ SGFR |
| VTFQ ↓ GKF                                | 3CLpro/TM3 | VTFQ ↓ SAVK |
| ATVQ ↓ SKMS                               | TM3/? | ATVQ ↓ SKMS |
| ATLQ ↓ AIAS                               | ? | ATLQ ↓ AIAS |
| VRLQ ↓ AGNA                               | GFL/? | VRLQ ↓ AGNA |
| PLMQ ↓ SADA                               | GFL/? | PLMQ ↓ SADA |
| TVLG ↓ AVGA                               | ?/RdRp | TVLG ↓ AVGA |
| ATLQ ↓ AENV                               | RdRp/NTPase | ATLQ ↓ AENV |
| TRLQ ↓ SLEN                               | NTPase, etc./exonuclease | TRLQ ↓ SLEN |
| PKLQ ↓ ASQA                               | exonuclease/2'-O-MT | PKLQ ↓ ASQA |

Table 2. Proposed polyprotein cleavage sites of Wuhan coronavirus compared to SARS CoV 3CL\textsuperscript{pro} (Table adapted from Pillaiyar et al).\textsuperscript{14} Bold residues represent differences; TM = Transmembrane; GFL = growth factor-like domain; RdRp = RNA-dependent RNA polymerase; 2'-O-MT, = 2'-O-methyltransferase

Methods.

The Wuhan coronavirus genome was obtained from Genbank/NCBI (release MN908947.1).\textsuperscript{4} Sequences from crystal structure sequences were obtained from the Protein Data Bank (rcsb.org). Sequence analysis was carried out using the web interface at the European Bioinformatics Institute (EMBL-EBI).\textsuperscript{18} Pairwise sequence alignments were performed using the EMBOSS-Needle method with default EBLOSUM62 matrices. Multiple sequence alignments were carried out using Clustal Omega using default HMM profile-profile parameters. Blast searches were carried out against the full UniProt Knowledgebase using blastp 2.9.0+ and default parameters.

Homology models were prepared using the publicly accessible online Swissmodel\textsuperscript{19-23} programs, searching the SWISS-MODEL template library (SMTL version 2020-01-02, and PDB release 2019-12-27), or using a directed user-defined template approach. Coronavirus crystal structures are predominantly homodimers and were modelled as such, but the further refinement was initially performed on the monomeric forms. Newer releases of the Swissmodel software enable the inclusion of bound inhibitors when appropriate. In this case, models derived from PDB:2a5i included a bound covalent inhibitor. The initial heavy atom-only models were further refined using the Protein Preparation module in Schrödinger Suite 2019-2\textsuperscript{24} to add hydrogen atoms and optimise the internal hydrogen bonding network. Finally, the models were energy minimised using the OPLS3e force field with charges from the force field and implicit water solvent, and the Polak-Ribier Conjugate Gradient (PRCG) method to gradient <0.05Å. The final minimised models were then analysed using MolProbity\textsuperscript{25} and visualised in Pymol v2.1.\textsuperscript{26} Molecular dynamics simulations were performed with the Desmond Molecular Dynamics System (D. E. Shaw Research, New York, NY) by using the tools incorporated in the Schrödinger Suite 2019-2.\textsuperscript{24}
Conclusion.

The main protease encoded by the Wuhan fish market coronavirus is very highly homologous to SARS CoV 3CL\textsuperscript{pro}, whereas the virus overall is more highly related to bat coronaviruses. Homology models of the Wuhan coronavirus protease were built from known SARS 3CL\textsuperscript{pro} crystal structures in both ligand-bound and apo forms. These represent potential starting points for structure-based drug design and for suggesting point mutations to inform future biochemical experiments.

Supporting Information

EMBOSS and Clustal Omega sequence alignments, MolProbity statistics and Ramachandran plots for refined monomer and dimer models. PDB files for raw Swissmodel and individual refined monomer and dimer models. Molecular Dynamics simulations of monomer and dimeric models.

References.

1. https://www.who.int/westernpacific/emergencies/pneumonia-in-wuhan-china. Accessed 12\textsuperscript{th} January 2020.

2. https://www.who.int/csr/don/12-january-2020-novel-coronavirus-china/en/. Accessed 12\textsuperscript{th} January 2020.

3. Holmes EC on behalf of Zhang Y-Z. http://virological.org/t/initial-genome-release-of-novel-coronavirus/319. Accessed 12\textsuperscript{th} January 2020.

4. Zhang Y-Z, Wu F, Chen Y-M, Pei Y-Y, Xu L, Wang W, Zhao S, Yu B, Hu Y, Tao Z-W, Song Z-G, Tian J-H, Zhang Y-L, Liu Y, Zheng J-J, Dai F-H, Wang Q-M, She J-L, and Zhu T-Y. Nucleotide Accession No. MN908947.1, Wuhan seafood market pneumonia virus isolate Wuhan-Hu-1, complete genome. Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information; [1988]. Accessed 13\textsuperscript{th} January 2020. Available from: https://www.ncbi.nlm.nih.gov/nuccore/MN908947

5. https://www.sciencemag.org/news/2020/01/mystery-virus-found-wuhan-resembles-bat-viruses-not-sars-chinese-scientist-says. Accessed 12\textsuperscript{th} January 2020.

6. Agbowuro AA, Huston WM, Gamble AB, Tyndall JDA. Proteases and protease inhibitors in infectious diseases. Med Res Rev. 2018;38(4): 1295-1331. (DOI:10.1002/med.21475)

7. Piot P, Abdool Karim SS, Hecht R, et al. Defeating AIDS--advancing global health. Lancet. 2015;386(9989): 171-218. (DOI:10.1016/s0140-6736(15)60658-4)
8. Howe AY, Venkatraman S. The Discovery and Development of Boceprevir: A Novel, First-generation Inhibitor of the Hepatitis C Virus NS3/4A Serine Protease. *J Clin Transl Hepatol.* 2013;1(1): 22-32. (DOI:10.14218/jeth.2013.002xx)

9. Drosten C, Gunther S, Preiser W, et al. Identification of a novel coronavirus in patients with severe acute respiratory syndrome. *N Engl J Med.* 2003;348(20): 1967-1976. (DOI:10.1056/NEJMoa030747)

10. Ksiazek TG, Erdman D, Goldsmith CS, et al. A novel coronavirus associated with severe acute respiratory syndrome. *N Engl J Med.* 2003;348(20): 1953-1966. (DOI:10.1056/NEJMoa030781)

11. Zaki AM, van Boheemen S, Bestebroer TM, Osterhaus AD, Fouchier RA. Isolation of a novel coronavirus from a man with pneumonia in Saudi Arabia. *N Engl J Med.* 2012;367(19): 1814-1820. (DOI:10.1056/NEJMoa1211721)

12. Ratia K, Saikatendu KS, Santarsiero BD, et al. Severe acute respiratory syndrome coronavirus papain-like protease: structure of a viral deubiquitinating enzyme. *Proc Natl Acad Sci U S A.* 2006;103(15): 5717-5722. (DOI:10.1073/pnas.0510851103)

13. Chen S, Chen L, Tan J, et al. Severe acute respiratory syndrome coronavirus 3C-like protease N terminus is indispensable for proteolytic activity but not for enzyme dimerization. Biochemical and thermodynamic investigation in conjunction with molecular dynamics simulations. *J Biol Chem.* 2005;280(1): 164-173. (DOI:10.1074/jbc.M408211200)

14. Pillaiyar T, Manickam M, Namavivayam V, Hayashi Y, Jung S-H. An Overview of Severe Acute Respiratory Syndrome–Coronavirus (SARS-CoV) 3CL Protease Inhibitors: Peptidomimetics and Small Molecule Chemotherapy. *Journal of Medicinal Chemistry.* 2016;59(14): 6595-6628. (DOI:10.1021/acs.jmedchem.5b01461)

15. Fan K, Wei P, Feng Q, et al. Biosynthesis, purification, and substrate specificity of severe acute respiratory syndrome coronavirus 3C-like protease. *J Biol Chem.* 2004;279(3): 1637-1642. (DOI:10.1074/jbc.M310875200)

16. 6 Bat coronavirus HKU4 crystal structures are available in the PDB; Accession codes 2YNA, 2YNB, 4YO9, 4YOG, 4YOI, 4YOJ.

17. St John SE, Tomar S, Stauffer SR, Mesecar AD. Targeting zoonotic viruses: Structure-based inhibition of the 3C-like protease from bat coronavirus HKU4--The likely reservoir host to the human coronavirus that causes Middle East Respiratory Syndrome (MERS). *Bioorg Med Chem.* 2015;23(17): 6036-6048. (DOI:10.1016/j.bmc.2015.06.039)
18. Madeira F, Park YM, Lee J, et al. The EMBL-EBI search and sequence analysis tools APIs in 2019. *Nucleic acids research*. 2019;47(W1): W636-W641. (DOI:10.1093/nar/gkz268)

19. Arnold K, Bordoli L, Kopp J, Schwede T. The SWISS-MODEL workspace: a web-based environment for protein structure homology modelling. *Bioinformatics*. 2006;22(2): 195-201. (DOI:10.1093/bioinformatics/bti770)

20. Kopp J, Schwede T. The SWISS-MODEL Repository: new features and functionalities. *Nucleic Acids Research*. 2006;34(suppl_1): D315-D318. (DOI:10.1093/nar/gkj056)

21. Biasini M, Bienert S, Waterhouse A, et al. SWISS-MODEL: modelling protein tertiary and quaternary structure using evolutionary information. *Nucleic Acids Res.* 2014;42(Web Server issue): W252-258. (DOI:10.1093/nar/gku340)

22. Bienert S, Waterhouse A, de Beer Tjaart AP, et al. The SWISS-MODEL Repository—new features and functionality. *Nucleic Acids Research*. 2017;45(D1): D313-D319. (DOI:10.1093/nar/gkw1132)

23. Waterhouse A, Bertoni M, Bienert S, et al. SWISS-MODEL: homology modelling of protein structures and complexes. *Nucleic Acids Research*. 2018;46(W1): W296-W303. (DOI:10.1093/nar/gky427)

24. Schrodinger L. Small-Molecule Drug Discovery Suite 2019-3. 2017-2 ed. New York, NY: Schrodinger, LLC; 2019.

25. Davis IW, Leaver-Fay A, Chen VB, et al. MolProbity: all-atom contacts and structure validation for proteins and nucleic acids. *Nucleic Acids Res.* 2007;35(Web Server issue): W375-383. (DOI:10.1093/nar/gkm216)

26. The PyMOL Molecular Graphics System, Version 2.0 Schrödinger, LLC.
Supporting Information for

Homology models of Wuhan Coronavirus 3CL\textsuperscript{pro} Protease

Dr. Martin J. Stoermer
Division of Chemistry and Structural Biology, Institute for Molecular Bioscience
The University of Queensland
St. Lucia, 4072, Queensland, Australia
E-mail: martin.stoermer@uq.edu.au
Orcid: https://orcid.org/0000-0003-3445-2104
Twitter: @MartinStoermer

Nomenclature

Original Swissmodel homology models were numbered 1-8. However due to redundancy monomer Models 1,2,4,6,7,8 were taken through to refinement. Model 8 was only obtained as a monomer, so just 5 models were refined as their homodimers. For clarity and to define their point of origin, these were named Models11,12,14,16,17&18.

E.g.:

| Raw Model | PDB Template | Refined monomer model | Refined dimer model |
|-----------|--------------|-----------------------|--------------------|
| 1 (dimer) | 2z9j         | Model 1               | Model 11           |
| 2 (dimer) | 3vb3         | Model 2               | Model 12           |
| 3 (dimer) | 2z9j         | Model 4               | Model 14           |
| 4 (dimer) | 1uk3         | Model 6               | Model 16           |
| 5 (dimer) | 1uk3         | Model 7               | Model 17           |
| 7 (dimer) | 1uj1         |                       |                    |
| 8 (monomer) | 1z1i       | Model 8               |                    |

Table S1. Naming system used for the models at each stage of refinement.
Figure S1. EMBOSS Needle alignment of the extracted protease sequence of Wuhan coronavirus against Bat coronavirus (PDB:2YNB) Catalytic dyads (His41, Cys145) shown in red bold.
# Aligned_sequences: 2
# 1: WH-Human_1pr
# 2: SARS_pr (2z9j_A)
# Matrix: EBLOSUM62
# Gap_penalty: 10.0
# Extend_penalty: 0.5

# Length: 306
# Identity: 294/306 (96.1%)
# Similarity: 302/306 (98.7%)
# Gaps: 0/306 (0.0%)
# Score: 1600.0

WH-Human_1pr
1 SGFRKMAFPSGKVEGCMVQVTCGGTTLNLGLWDVVYCPRHICTSEDML 50
SARS_pr
1 SGFRKMAFPSGKVEGCMVQVTCGGTTLNLGLWDVVYCPRHICTAEML 50

WH-Human_1pr
51 NPNYEDLLIRKSNHNFVQAGNVQRVIGHSMQNCVLKLVDVTANPKTPK 100
SARS_pr
51 NPNYEDLLIRKSNHNFVQAGNVQRVIGHSMQNCVLKLVDVTANPKTPK 100

WH-Human_1pr
101 YKFVRIQPGQTFSVLACYNGSPSGVYQCAMRPNFTIKGSLNGSCGSVGF 150
SARS_pr
101 YKFVRIQPGQTFSVLACYNGSPSGVYQCAMRPNFTIKGSLNGSCGSVGF 150

WH-Human_1pr
151 NIDYDCVSFCYMHHMELPTGVHAGTDLEGNYGFVDPQTAQAGTDTTI 200
SARS_pr
151 NIDYDCVSFCYMHHMELPTGVHAGTDLEGNKGYGFVDPQTAQAGTDTTI 200

WH-Human_1pr
201 TYNVLWLYAVINGDRWFLNFRTTTLNDFNVLAMKYNP3LQDHVDIL 250
SARS_pr
201 TYNVLWLYAVINGDRWFLNFRRTTLNDFNVLAMKYNP3LQDHVDIL 250

WH-Human_1pr
251 GPLSAQTGIAVLDMCAALKEKLLQNGMNGRTELGSALLEDEFTPVVRQC 300
SARS_pr
251 GPLSAQTGIAVLDMCAALKEKLLQNGMNGRTELGSALLEDEFTPVVRQC 300

WH-Human_1pr
301 SGVTFRQ 306
SARS_pr
301 SGVTFRQ 306

Figure S2. EMBOSS Needle alignment of the extracted protease sequence of Wuhan coronavirus against SARS coronavirus (PDB: 2z9j) Catalytic dyads shown in red bold.

MolProbity scoring and statistics for refined monomer models 1,2,4,6,7,8 and dimer models

Monomer models.

| Model | Template name | Ramachandran favoured (%) | Ramachandran allowed (%) | Clashscore (percentile) | MolProbity score (percentile) | CA Geometry outliers (%) | CAβ deviations (%) |
|-------|---------------|---------------------------|--------------------------|-------------------------|-------------------------------|--------------------------|-------------------|
| 1     | 2z9j.1.B      | 96.33                     | 100                      | 1.52 (99⁰)              | 1.36 (98⁰)                    | 0.0                      | 0.0               |
| 2     | 3vb3.1.A      | 96.66                     | 100                      | 1.30 (99⁰)              | 1.12 (100⁰)                  | 0.0                      | 0.0               |
| 4     | 1nk3.1.B      | 93.98                     | 99.67                    | 1.52 (99⁰)              | 1.74 (88⁰)                    | 0                        | 0                 |
| 6     | 2a51.1.A      | 96.38                     | 99.67                    | 0.63 (99⁰)              | 0.95 (100⁰)                  | 0                        | 0                 |
| 7     | 1ujl.1.B      | 96.66                     | 99.67                    | 1.52 (99⁰)              | 1.43 (96⁰)                    | 0                        | 0                 |
| 8     | 1zl1.1.A      | 91.30                     | 99.33                    | 3.04 (98⁰)              | 2.02 (75⁰)                    | 0.34                     | 0.36              |

Table S2. Statistics via MolProbity with no His/Asn/Gln flips allowed.
Table S3. Statistics via MolProbity with His/Asn/Gln flips allowed. 5/6 models had the possibility of suggested hydrogen bonding flips, but only the top model1 resulted in any improvement to the clash and MolProbity scores.

**Dimer models:**

| Dimer | Template name | Ramachandran favoured (%) | Ramachandran allowed (%) | Clashscore (percentile) | MolProbity score (percentile) | CA Geometry outliers (%) | Cβ deviations (%) |
|-------|---------------|---------------------------|--------------------------|--------------------------|------------------------------|-------------------------|------------------|
| 11    | 2v9j.1.B      | 96.51                     | 99.83                    | 1.62 (99th)              | 1.41 (97th)                  | 0.17                    | 0.0              |
| 12    | 3vb3.1.A      | 96.35                     | 100                      | 1.18 (99th)              | 1.12 (100th)                 | 0.0                     | 0.0              |
| 14    | 1uk3.1.B      | 94.16                     | 99.67                    | 1.84 (99th)              | 1.73 (88th)                  | 0.34                    | 0.0              |
| 16    | 2a5i.1.A      | 95.72                     | 99.67                    | 0.73 (99th)              | 1.03 (100th)                 | 0                       | 0.0              |
| 17    | 1uj1.1.B      | 94.82                     | 99.83                    | 1.30 (99th)              | 1.58 (93rd)                  | 0.34                    | 0.18             |

Table S4. Statistics via MolProbity with no His/Asn/Gln flips allowed.

| Dimer | Template name | Ramachandran favoured (%) | Ramachandran allowed (%) | Clashscore (percentile) | MolProbity score (percentile) | CA Geometry outliers (%) | Cβ deviations (%) |
|-------|---------------|---------------------------|--------------------------|--------------------------|------------------------------|-------------------------|------------------|
| 11 (2 flips) | 2v9j.1.B      | 96.51                     | 99.83                    | 1.51 (99th)              | 1.40 (97th)                  | 0.17                    | 0.0              |
| 12 (4 flips) | 3vb3.1.A      | 96.35                     | 100                      | 1.18 (99th)              | 1.12 (100th)                 | 0.0                     | 0.0              |
| 14 (1 flip)   | 1uk3.1.B      | 94.16                     | 99.67                    | 1.84 (99th)              | 1.73 (88th)                  | 0.34                    | 0.0              |
| 16 (4 flips)  | 2a5i.1.A      | 95.72                     | 99.67                    | 0.73 (99th)              | 1.03 (100th)                 | 0                       | 0.0              |
| 17 (3 flips)  | 1uj1.1.B      | 94.82                     | 99.83                    | 1.30 (99th)              | 1.58 (93rd)                  | 0.34                    | 0.18             |

Table S5. Statistics via MolProbity with His/Asn/Gln flips allowed.
MolProbity Ramachandran analysis

WH_human_protease_from2z9j_Minimised1H.pdb, model 1

96.3% (289/300) of all residues were in favored (98%) regions.
100.0% (300/300) of all residues were in allowed (>99.8%) regions.

Figure S2. Ramachandran plots for Wuhan CoV 3CL\textsuperscript{pro} homology monomer model1
MolProbity Ramachandran analysis

WH_human_protease_from3vb3_Minimised1H.pdb, model 2

96.7% (289/299) of all residues were in favored (98%) regions.
100.0% (299/299) of all residues were in allowed (>99.8%) regions.

Figure S3. Ramachandran plots for Wuhan CoV 3CL\textsuperscript{Pro} homology monomer model 2
MolProbity Ramachandran analysis

WH_human_protease_from1uk3_Minimised1H.pdb, model 4

94.0% (281/299) of all residues were in favored (98%) regions.
99.7% (298/299) of all residues were in allowed (>99.8%) regions.

Figure S4. Ramachandran plots for Wuhan CoV 3CLpro homology monomer model 4
96.4% (293/304) of all residues were in favored (98%) regions.
99.7% (303/304) of all residues were in allowed (>99.8%) regions.

Figure S5. Ramachandran plots for Wuhan CoV 3CL\textsuperscript{pro} homology monomer model6
MolProbity Ramachandran analysis

WH_human_protease_from1uj1_Minimised1H.pdb, model 7

96.7% (289/299) of all residues were in favored (98%) regions.
99.7% (298/299) of all residues were in allowed (>99.8%) regions.

Figure S6. Ramachandran plots for Wuhan CoV 3CLpro homology monomer model 7
MolProbity Ramachandran analysis

WH_human_protease_from1z1i_Minimised1H.pdb, model 8

91.3% (273/299) of all residues were in favored (98%) regions.
99.3% (297/299) of all residues were in allowed (>99.8%) regions.

Figure S7. Ramachandran plots for Wuhan CoV 3CL<sup>pro</sup> homology monomer model8
96.5% (581/602) of all residues were in favored (98%) regions.
99.8% (601/602) of all residues were in allowed (>99.8%) regions.
**Figure S8.** Ramachandran plots for Wuhan CoV 3CL\textsuperscript{pro} homology dimer model11
MolProbity Ramachandran analysis

WH_human_protease_from3vb3_dimerMinimised1H.pdb, model 12

96.4% (581/603) of all residues were in favored (98%) regions.
100.0% (603/603) of all residues were in allowed (>99.8%) regions.

**Figure S9.** Ramachandran plots for Wuhan CoV 3CL\textsuperscript{pro} homology dimer model12
MolProbity Ramachandran analysis

WH_human_protease_from1uk3_dimerMinimised1H.pdb, model 14

94.2% (564/599) of all residues were in favored (98%) regions.
99.7% (597/599) of all residues were in allowed (>99.8%) regions.

Figure S10. Ramachandran plots for Wuhan CoV 3CLpro homology dimer model14
MolProbity Ramachandran analysis

WH_human_protease_from2a5i_dimerMinimised1H.pdb, model 16

95.7% (582/608) of all residues were in favored (98%) regions.
99.7% (606/608) of all residues were in allowed (>99.8%) regions.

Figure S11. Ramachandran plots for Wuhan CoV 3CLpro homology dimer model16
MolProbity Ramachandran analysis

WH_human_protease_from1uj1_dimerMinimised1H.pdb, model 17

94.8% (568/599) of all residues were in favored (98%) regions.
99.8% (598/599) of all residues were in allowed (>99.8%) regions.

Figure S12. Ramachandran plots for Wuhan CoV 3CLpro homology dimer model17
Molecular Dynamics Method

Molecular dynamics simulations were performed with the Desmond Molecular Dynamics System (D. E. Shaw Research, New York, NY) by using the tools incorporated in the Schrödinger Suite 2019-2. The model system was set up by using the “System Setup” utility. The OPLS2005 force field parameters were implemented during simulations. Each monomer, dimer or ternary complex was placed in an orthorhombic box with solvent buffers up to 10Å along each dimension. The TIP3P solvent model was used to describe water molecules. Overall neutralisation of the system was achieved by adding sodium and chloride ions at the physiological concentration of 0.15M. The prepared model systems were then relaxed by using the multi-step default protocol in Desmond. After relaxation, the whole system was subjected to a 100ns simulation. The cut-off distance for computing short-range electrostatics and Lennard–Jones interaction was set to 9.0Å. The trajectories and energies were recorded every 100 and 1.2ps, respectively. MD simulations were performed at the IMB cluster from the University of Queensland on NVIDIA P100 GPUs or on a CentOS Linux Desktop with a Quadro RTX 5000 GPU. Simulation analysis was performed on Linux or Macintosh desktops using the Simulation Interactions Diagrams module in Schrödinger Suite 2019-2.

Model2 from SARS 3vb3 WH_human_protease_from3vb3_Minimised.pdb

Figure S1-13. Protein C-alpha RMSD over the 100 ns duration molecular dynamics simulation. Figure S1-14. Protein C-alpha RMSF plot. Per residue variability in Cα position.

Flexible residues

Figure S1-15. WH CoV 3CLpro Model 2 shown as green ribbons, except catalytic dyad (His41, Cys145 – yellow sticks), and magenta ribbons for regions found to be mobile in 100 ns MD simulation.
Moderately mobile regions (>1.5 RMS) N-terminal residues Ser1-Phe3, β-hairpin loop Gly23-Thr25, α-helix and loop Thr45-Arg60, loop tip Asn72, strand Gly138-Gly143, β-hairpin loop Pro168-Gly170, long strand Gln189-Gly195, strand Asn274-Gly278, C-terminus Gln299-Ser301.

*Model6 from SARS 2a5i WH_human_protease_from2a5i_Minimised.pdb*

**Figure S1-16.** Protein C-alpha (blue line) and ligand (red line) RMSD over the 100 ns duration MD run. **Figure S1-17** Protein RMSF plot. Per residue variability in Cα position. **Figure S1-18.** Protein- ligand interaction summary diagram. Orange: negative charge interaction, blue: polar interactions, green: hydrophobic interactions, grey circles: solvent exposed atoms.

Over the first 10 ns of the simulation the mobile loop 44CTSEDMLNPN53 defining the boundary of the S2 subsite moves away from the bound ligand, and remains distant during the remaining 80ns.
Flexible residues

**Figure S1-19.** WH CoV 3CL\(^{pro}\) Model 6 shown as green ribbons, except catalytic dyad (His41, Cys145 – yellow sticks), and magenta ribbons for regions found to be mobile in the 100 ns molecular dynamics simulation. Covalently bound SARS inhibitor carried through in model from template PDB:2a5i shown as grey sticks.

Moderately mobile regions (>1.5 RMSD) N-terminal residues Ser1-Gly2, \(\beta\)-hairpin loop Gly23-Thr25, very mobile \(\alpha\)-helix and loop Thr45-Arg60, loop tip Asn72, long strand Arg188-Ala194, strand Arg222-Thr224, strand Asn274-Arg279, C-terminus Ser301-Gln306.

*Model12 from SARS 3vb3 WH_human_protease_from3vb3_dimerMinimised.pdb*

**Figure S1-20.** Protein C-alpha (blue line) RMSD over the 100 ns duration MD run. **Figure S1-21.** Protein RMSF plot. Per residue variability in C\(\alpha\) position.
Flexible residues

**Figure S1-21.** WH CoV 3CL\(^{pro}\) Model 12 Chain A shown as green ribbons, Chain B cyan ribbons, except catalytic dyads (His41, Cys145 – green and cyan sticks), and magenta ribbons for regions found to be mobile in 100 ns MD simulation.

In comparison to the MD simulation of the monomer (Model 2), the dimer shows stable N- and C-termini for both subunits, as these form part of the dimer interface which remains quite stable during the simulation. The very mobile helix and loop region show slight differences in mobility (Chain A Ser46-Ser62, Chain B Thr45-Asp56). In addition there are differences in the N-terminal domains, the Chain B helices Leu227-Met235 and Glu270-Met276 being slightly more mobile than those in Chain A.

Moderately mobile regions (>1.5 RMSD) Chain A: \(\alpha\)-helix and loop Ser46-Ser62, loop tip Asn72, long strand 215-Thr225, helical turn Thr243-His246, large loop Asn274-Arg279, C-terminus Ser301, Chain B: \(\beta\)-hairpin loop Gly23-Thr25, \(\alpha\)-helix and loop Thr45-Asp56, loop tip Asn72, loop Pro168-Gly170, long strand and helix Trp218-Met235, loop Thr243-Gln244, loop and helical turn Glu270-Thr280, loop Gly283-Leu286.

**Model16 from SARS 2a5i WH_human_protease_from2a5i_dimerMinimised.pdb**

**Figure S1-22.** Protein C-alpha (blue line) and ligand (red line) RMSD over the 100 ns duration MD run. **Figure S1-23.** Protein RMSF plot. Per residue variability in C\(\alpha\) position.
Figure S1-24. Protein-ligand interaction summary diagram. Orange: negative charge interaction, blue: polar interactions, green: hydrophobic interactions, grey circles: solvent exposed atoms.

Flexible residues

Figure S1-14. WH CoV 3CL\textsuperscript{pro} Model 16 Chain A shown as green ribbons, Chain B cyan ribbons, except catalytic dyads (His41, Cys145 – yellow sticks), and magenta ribbons for regions found to be mobile in 100 ns MD simulation. SARS 3CL\textsuperscript{pro} inhibitor shown as grey sticks.
Apart from the more stable N- and C-termini for both subunits which form part of the dimer interface the remainder of the dimer remains of similar overall flexibility. One major difference is that in the dimeric form the loop $^{44}$CTEDMLNP$^{53}$ which forms the boundary of the S2-subsite which binds bulky substrate residues Met/Phe/Leu is much less mobile than was seen in the 100ns simulation of the monomer.

Moderately mobile regions (>1.5 RMSD) Chain A: loop Thr45-Asp48, $\alpha$-helix Asn53-Arg60, loop tip Asn72, loop tip Tyr154, strand Ala191-Ala193, strand residue Phe223, strand Asn274-Gly278, C-terminus Gly302-Gln306. Chain B: $\beta$-hairpin loop Gly23-Thr24, large loop Cys44-Asp53, strand Ile59-His64, loop tip Asn72, loop tip Tyr154, long strand and loop Val186-Gly195, strand residue Gly215, strand Phe223-Thr224, $\alpha$-helical turn Leu232-Tyr237, helix residue Val247, helical turn Ala255-Gly258, very mobile loop and helical turn Glu270-Leu286, C-terminus Gly302-Gln306
Homology models of Wuhan Coronavirus 3CL<sup>pro</sup> Protease

Martin J. Stoermer*
Division of Chemistry and Structural Biology, Institute for Molecular Bioscience, The University of Queensland, St. Lucia, 4072, Queensland, Australia.

*Corresponding author
Dr. Martin J. Stoermer
Division of Chemistry and Structural Biology, Institute for Molecular Bioscience
The University of Queensland
St. Lucia, 4072, Queensland, Australia
E-mail: martin.stoermer@uq.edu.au
Orcid: https://orcid.org/0000-0003-3445-2104
Twitter: @MartinStoermer

Abstract.

A regional outbreak of pneumonia in Wuhan, Hubei Province of China in late 2019 was associated with a novel coronavirus. Rapid release of genomic data for the isolated virus enabled the construction of first-generation homology models of the new CoV 3CL<sup>pro</sup> cysteine protease. Whilst the overall viral genome was most closely associated with bat coronaviruses, the main protease is most closely related (96% identity) to SARS CoV protease.

Keywords: Wuhan coronavirus, 3CL<sup>pro</sup>, protease, homology modelling
Foreword and Acknowledgements

The author is indebted to Professor Yong-Zhen Zhang, Fudan University, Shanghai, the Shanghai Public Health Clinical Center & School of Public Health, the Central Hospital of Wuhan, Huazhong University of Science and Technology, the Wuhan Center for Disease Control and Prevention, the National Institute for Communicable Disease Control and Prevention, Chinese Center for Disease Control, and Professor Edward Holmes, The University of Sydney, Sydney, Australia for the rapid provision of the Wuhan coronavirus genome sequence WH-Human_1 and subsequent GenBank submissions to the international research community and the public.

Introduction.

In December 2019, the World Health Organisation (WHO) was officially informed of a cluster of cases of pneumonia of unknown cause detected in Wuhan City, Hubei Province of China.\(^1\) Subsequently researchers from the Shanghai Public Health Clinical Center & School of Public Health reported that the evidence suggested an origin in a seafood market in Wuhan, and the isolation of a new type of coronavirus (novel coronavirus, nCoV) on 7 January 2020. Laboratory testing ruled out other respiratory pathogens including Severe Acute Respiratory Syndrome coronavirus (SARS-CoV).\(^2\)

On 10th January 2020 a preliminary sequence (WH-Human_1.fasta.gz) of the novel coronavirus was made available on virological.org,\(^3\) and subsequently on 12\(^{th}\) January 2020, was released on Genbank (ID MN908947).\(^4\) Preliminary analysis of the Wuhan virus suggests it is most closely related to coronaviruses found in bats.

Proteases are an important class of drug target,\(^5\) most successfully in the cases of HIV protease\(^7\) and Hepatitis C\(^8\). Recent human outbreaks of related coronaviruses SARS\(^9,10\) in 2002 and MERS\(^11\) in 2012 have focused the world’s attention on the global risks of such emerging viruses. Coronaviruses are positive strand, enveloped RNA viruses with exceptionally large genomes which usually encode two or three viral proteases. In the case of Severe Acute Respiratory Syndrome coronavirus, the two proteases are a papain-like cysteine protease (PL\(^{pro}\))\(^12\) and a chymotrypsin-like cysteine protease 3CL\(^{pro}\),\(^13\) also known as the Main protease. SARS-CoV 3CL\(^{pro}\) is essential for viral replication and is thus recognized as a potential drug target for the treatment of SARS infections. To date there are no clinically used inhibitors of the SARS protease yet they remain in development.\(^14\)
Figure 1. X-ray crystal structure of SARS-CoV 3CLpro bound to a covalent inhibitor (PDB code 2a5i). Panel A: Domain I shown as yellow ribbons, domain II cyan ribbons, domain III pale pink ribbons. Catalytic dyad His41, Cys145 shown as yellow sticks, covalent azapeptide inhibitor shown as grey sticks. Panel B: Active site at the interface of domains I, II with P1 Gln residue shown hydrogen bonding to S1 subsite residues Phe140, His163, and Glu166 (blue lines).

The active site of SARS-CoV 3CLpro is located in the cleft between domains I and II of the protein, and the two domains each contribute one residue to the catalytic dyad composed of His41 and Cys145 (Figure 1). Substrates and inhibitors typically contain Gln residues at the P1 position and large hydrophobic residues Phe/Leu/Met at P2.15 The wealth of medicinal chemistry and biochemical information gained over the last 18 years, coupled with good-high sequence similarity between the members of the coronavirus family makes the homology modelling of new and emerging viruses an inviting prospect.

Results and Discussion.

The release of the viral genome of the novel Wuhan coronavirus prompted us to construct first-generation homology models of the new CoV 3CLpro. Initial studies and BLAST analysis indicated that the virus was most closely related to a clade of bat coronaviruses (Figure 2), of which the closest neighbour for which there was protease structural information available in the Protein Data Bank (www.rcsb.org) was bat coronavirus HKU4.16,17 EMBOSS Needle18 alignment of the Wuhan genome with the sequence extracted from PDB:2YNB identified a putative 306ss viral 3CLpro protein with 49% sequence identity (65% sequence similarity) (Figure 3).
Figure 2. Phylogenetic Neighbour-joining tree without distance corrections for Wuhan coronavirus and top BLAST hits generated with Clustal Omega.
An initial Swissmodel\textsuperscript{19} search for possible homology model templates against this search sequence however provided only SARS protease templates with much higher sequence identity and similarity (96, 99\% respectively). This was confirmed via a second EMBOSS Needle alignment (Figure 3), and visually by a mapping of differing residues onto the respective crystal structures (Figure 4).

**Figure 3.** EMBOSS Needle alignment of the extracted protease sequence of Wuhan coronavirus against Bat coronavirus (PDB: 2ynb) and SARS (PDB: 2z9j). Catalytic dyads shown in red bold.

**Figure 4.** Visualisation of sequence differences between bat HKU4, SARS and Wuhan coronaviruses. Panel A: X-ray crystal structure (PDB: 2z9j) of SARS CoV protease shown as green ribbons, with catalytic His41 and Cys145 residues shown as cyan sticks. Residues which differ in the Wuhan sequence are marked in magenta. Panel B: X-ray crystal structure
(PDB:2ynb) of bat coronavirus HKU4 protease shown as green ribbons, with residues which differ in the Wuhan sequence are marked in magenta.

Alignment of the SARS 2z9j and HKU4 2ynb protease crystal structures however showed that the two had a very similar overall fold with an RMSD of 0.66Å across 229 well conserved Cα atoms. Accordingly, due to their higher overall sequence identities, SARS crystal structures were chosen for the future development of Wuhan CoV protease homology models. Coronavirus crystal structures are predominantly homodimers, and as a consequence Swissmodel template searches returned some duplicates arising from both chains A&B being selected. Models were built as homodimers where appropriate however for visualisation and initial refinement purposes the monomers were used (Table 1, Figure 5A). Pairs 1&3, 4&5 were functionally identical so only 6 models were taken through for further refinement. The 5 Swissmodel dimer models 1,2,4,6,7 were refined as both dimers and monomers, using Schrodinger Suite 2019 and monomer model 8 was refined without reconstruction to a dimer. The models were overlaid as their monomers and compared by manual inspection. The models were very structurally similar, especially in the active site region near His41 and Cys145 (Figure 5B).

| Mode | Template name | Title          | % Identity | QSQE score | Method (Resolution) | Oligo State | Ligand |
|------|---------------|----------------|------------|------------|---------------------|-------------|--------|
| 1    | 2z9j.1.B      | 3C-like proteinase | 96.1       | 0.91       | X-ray, 2.0Å         | homodimer   | 2x DTZ |
| 2    | 3vb3.1.A      | 3C-like proteinase | 96.1       | 0.91       | X-ray, 2.2Å         | homodimer   | None   |
| 3    | 2z9j.1.A      | 3C-like proteinase | 96.1       | 0.91       | X-ray, 2.0Å         | homodimer   | 2x DTZ |
| 4    | 1uk3.1.B      | 3C-like proteinase | 96.1       | 0.88       | X-ray, 2.4Å         | homodimer   | None   |
| 5    | 1uk3.1.A      | 3C-like proteinase | 96.1       | 0.88       | X-ray, 2.4Å         | homodimer   | None   |
| 6    | 2a5i.1.A      | 3C-like peptidase | 96.1       | 0.86       | X-ray, 1.9Å         | homodimer   | 2x AZP |
| 7    | 1uj1.1.B      | 3C-like proteinase | 96.1       | 0.86       | X-ray, 1.9Å         | homodimer   | None   |
| 8    | 1zli.1.A      | 3C-like proteinase | 96.1       | -          | X-ray, 2.8Å         | monomer     | None   |

Table 1. PDB templates used for Swissmodel homology model building. Templates ranked by internal Swissmodel QSQE score
By a combination of Swissmodel scoring of the initial models and MolProbity analysis of the refined overall model quality, the final two models selected were Model2 from PDB:3vb3 representing the unbound or apo form of the protease, and Model6 from PDB:2a5i which represents a ligand-bound form. The original SARS CoV 3CL\textsuperscript{pro} inhibitor was carried through the modelling protocol and is shown in the active site of the Wuhan 3CL\textsuperscript{pro} model (Figure 5). The apo-like and bound-like models 2,4 are very similar overall except for a significant difference in the loop 45TSEDM\textsuperscript{49} which forms the boundary of the S2 subsite. This results in the bulky Met49 side chain moving out of the S2 pocket allowing the ligand Phe residue to be accommodated.

These models were also analysed using 100ns molecular dynamics simulations (see Supporting Information) to check the systems for overall structural stability (Figure 6). In general the dimeric forms were slightly more stable overall, as the C- and N-termini of the
chains form part of the dimer interface and are held more closely than in the monomeric state where they are more mobile.

**Figure 6.** Comparison of mobile regions in molecular dynamics simulations. Panel A: Monomeric model2 as green ribbons except magenta ribbons for moderately mobile regions. Catalytic dyad yellow sticks. Panel B: Dimeric model12 shown as green ribbons Chain A, cyan ribbons Chain B, except magenta ribbons for moderately mobile regions. Catalytic dyad green and cyan sticks.

Finally an analysis of the new Wuhan coronavirus genome revealed multiple sites, analogous to those found or proposed for SARS\textsuperscript{14} that may be cleaved by the 3CL\textsubscript{pro} after (L/F/M)-Gln residues. These are mostly identical to the SARS sequences (Table 2).

| SARS 3CL\textsubscript{pro} cleavage site | Site | Potential Wuhan coronavirus sites |
|------------------------------------------|------|----------------------------------|
| AVLQ;SGFR                               | TM2/3CL\textsubscript{pro} | AVLQ;SGFR                      |
| VTFQ;GFK                                | 3CL\textsubscript{pro}/TM3 | VTFQ;SAVK                      |
| ATVQ;SKMS                                | TM3/? | ATVQ;SKMS                      |
| ATLQ;AIA5                                | ?    | ATLQ;AIA5                      |
| VKLQ;NNEL                                | ?    | VKLQ;NNEL                      |
| VRLQ;AGNA                                | ?/GFL | VRLQ;AGNA                      |
| PLMQ;SADA                                | GFL/? | PMLQ;SADA                      |
| TVLQ;AVGA                                | ?/RdRp | TVLQ;AVGA                      |
| ATLQ;AENV                                | RdRp/NTPase | ATLQ;AENV            |
| TRLQ;SLEN                                | NTPase, etc./exonuclease | TRLQ;SLEN                      |
| PKLQ;ASQA                                | exonuclease/2'-O-MT | PKLQ;ASQA                      |

**Table 2.** Proposed polyprotein cleavage sites of Wuhan coronavirus compared to SARS CoV 3CL\textsubscript{pro} (Table adapted from Pillaiyar et al.).\textsuperscript{14} Bold residues represent differences; TM = Transmembrane; GFL = growth factor-like domain; RdRp = RNA- dependent RNA polymerase; 2’-O-MT, = 2’-O-methyltransferase
Methods.

The Wuhan coronavirus genome was obtained from Genbank/NCBI (release MN908947.1). Sequences from crystal structure sequences were obtained from the Protein Data Bank (rcsb.org). Sequence analysis was carried out using the web interface at the European Bioinformatics Institute (EMBL-EBI). Pairwise sequence alignments were performed using the EMBOSS-Needle method with default EBLOSUM62 matrices. Multiple sequence alignments were carried out using Clustal Omega using default HMM profile-profile parameters. Blast searches were carried out against the full UniProt Knowledgebase using blastp 2.9.0+ and default parameters.

Homology models were prepared using the publicly accessible online Swissmodel programs, searching the SWISS-MODEL template library (SMTL version 2020-01-02, and PDB release 2019-12-27), or using a directed user-defined template approach. Coronavirus crystal structures are predominantly homodimers and were modelled as such, but the further refinement was initially performed on the monomeric forms. Newer releases of the Swissmodel software enable the inclusion of bound inhibitors when appropriate. In this case, models derived from PDB:2a5i included a bound covalent inhibitor. The initial heavy atom-only models were further refined using the Protein Preparation module in Schrödinger Suite 2019-2 to add hydrogen atoms and optimise the internal hydrogen bonding network. Finally, the models were energy minimised using the OPLS3e force field with charges from the force field and implicit water solvent, and the Polak-Ribier Conjugate Gradient (PRCG) method to gradient <0.05Å. The final minimised models were then analysed using MolProbity and visualised in Pymol v2.1. Molecular dynamics simulations were performed with the Desmond Molecular Dynamics System (D. E. Shaw Research, New York, NY) by using the tools incorporated in the Schrödinger Suite 2019-2.

Conclusion.

The main protease encoded by the Wuhan fish market coronavirus is very highly homologous to SARS CoV 3CL\textsubscript{pro}, whereas the virus overall is more highly related to bat coronaviruses. Homology models of the Wuhan coronavirus protease were built from known SARS 3CL\textsubscript{pro} crystal structures in both ligand-bound and apo forms. These represent potential starting points for structure-based drug design and for suggesting point mutations to inform future biochemical experiments.

Supporting Information
EMBOSS and Clustal Omega sequence alignments, MolProbity statistics and Ramachandran plots for refined monomer and dimer models. PDB files for raw Swissmodel and individual refined monomer and dimer models. Molecular Dynamics simulations of monomer and dimeric models.

References.

1. https://www.who.int/westernpacific/emergencies/pneumonia-in-wuhan-china. Accessed 12th January 2020.

2. https://www.who.int/csr/don/12-january-2020-novel-coronavirus-china/en/. Accessed 12th January 2020.

3. Holmes EC on behalf of Zhang Y-Z. http://virological.org/t/initial-genome-release-of-novel-coronavirus/319. Accessed 12th January 2020.

4. Zhang Y-Z, Wu F, Chen Y-M, Pei Y-Y, Xu L, Wang W, Zhao S, Yu B, Hu Y, Tao Z-W, Song Z-G, Tian J-H, Zhang Y-L, Liu Y, Zheng J-J, Dai F-H, Wang Q-M, She J-L, and Zhu T-Y. Nucleotide Accession No. MN908947.1, Wuhan seafood market pneumonia virus isolate Wuhan-Hu-1, complete genome. Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information; [1988]. Accessed 13th January 2020. Available from: https://www.ncbi.nlm.nih.gov/nuccore/MN908947

5. https://www.sciencemag.org/news/2020/01/mystery-virus-found-wuhan-resembles-bat-viruses-not-sars-chinese-scientist-says. Accessed 12th January 2020.

6. Agbowuro AA, Huston WM, Gamble AB, Tyndall JDA. Proteases and protease inhibitors in infectious diseases. Med Res Rev. 2018;38(4): 1295-1331. (DOI:10.1002/med.21475)

7. Piot P, Abdool Karim SS, Hecht R, et al. Defeating AIDS--advancing global health. Lancet. 2015;386(9989): 171-218. (DOI:10.1016/s0140-6736(15)60658-4)

8. Howe AY, Venkatraman S. The Discovery and Development of Boceprevir: A Novel, First-generation Inhibitor of the Hepatitis C Virus NS3/4A Serine Protease. J Clin Transl Hepatol. 2013;1(1): 22-32. (DOI:10.14218/jcth.2013.002xx)
9. Drosten C, Gunther S, Preiser W, et al. Identification of a novel coronavirus in patients with severe acute respiratory syndrome. *N Engl J Med.* 2003;348(20): 1967-1976. (DOI:10.1056/NEJMoa030747)

10. Ksiazek TG, Erdman D, Goldsmith CS, et al. A novel coronavirus associated with severe acute respiratory syndrome. *N Engl J Med.* 2003;348(20): 1953-1966. (DOI:10.1056/NEJMoa030781)

11. Zaki AM, van Boheemen S, Bestebroer TM, Osterhaus AD, Fouchier RA. Isolation of a novel coronavirus from a man with pneumonia in Saudi Arabia. *N Engl J Med.* 2012;367(19): 1814-1820. (DOI:10.1056/NEJMoa1211721)

12. Ratia K, Saikatendu KS, Santarsiero BD, et al. Severe acute respiratory syndrome coronavirus papain-like protease: structure of a viral deubiquitinating enzyme. *Proc Natl Acad Sci U S A.* 2006;103(15): 5717-5722. (DOI:10.1073/pnas.0510851103)

13. Chen S, Chen L, Tan J, et al. Severe acute respiratory syndrome coronavirus 3C-like proteinase N terminus is indispensable for proteolytic activity but not for enzyme dimerization. Biochemical and thermodynamic investigation in conjunction with molecular dynamics simulations. *J Biol Chem.* 2005;280(1): 164-173. (DOI:10.1074/jbc.M408211200)

14. Pillaiyar T, Manickam M, Namasivayam V, Hayashi Y, Jung S-H. An Overview of Severe Acute Respiratory Syndrome–Coronavirus (SARS-CoV) 3CL Protease Inhibitors: Peptidomimetics and Small Molecule Chemotherapy. *Journal of Medicinal Chemistry.* 2016;59(14): 6595-6628. (DOI:10.1021/acs.jmedchem.5b01461)

15. Fan K, Wei P, Feng Q, et al. Biosynthesis, purification, and substrate specificity of severe acute respiratory syndrome coronavirus 3C-like proteinase. *J Biol Chem.* 2004;279(3): 1637-1642. (DOI:10.1074/jbc.M310875200)

16. 6 Bat coronavirus HKU4 crystal structures are available in the PDB; Accession codes 2YNA, 2YNB, 4YO9, 4YOG, 4YOI, 4YOJ.

17. St John SE, Tomar S, Stauffer SR, Mesecar AD. Targeting zoonotic viruses: Structure-based inhibition of the 3C-like protease from bat coronavirus HKU4--The likely reservoir host to the human coronavirus that causes Middle East Respiratory Syndrome (MERS). *Bioorg Med Chem.* 2015;23(17): 6036-6048. (DOI:10.1016/j.bmc.2015.06.039)
18. Madeira F, Park YM, Lee J, et al. The EMBL-EBI search and sequence analysis tools APIs in 2019. *Nucleic acids research.* 2019;47(W1): W636-W641. (DOI:10.1093/nar/gkz268)

19. Arnold K, Bordoli L, Kopp J, Schwede T. The SWISS-MODEL workspace: a web-based environment for protein structure homology modelling. *Bioinformatics.* 2006;22(2): 195-201. (DOI:10.1093/bioinformatics/bti770)

20. Kopp J, Schwede T. The SWISS-MODEL Repository: new features and functionalities. *Nucleic Acids Research.* 2006;34(suppl_1): D315-D318. (DOI:10.1093/nar/gkj056)

21. Biasini M, Bienert S, Waterhouse A, et al. SWISS-MODEL: modelling protein tertiary and quaternary structure using evolutionary information. *Nucleic Acids Res.* 2014;42(Web Server issue): W252-258. (DOI:10.1093/nar/gku340)

22. Bienert S, Waterhouse A, de Beer Tjaart AP, et al. The SWISS-MODEL Repository—new features and functionality. *Nucleic Acids Research.* 2017;45(D1): D313-D319. (DOI:10.1093/nar/gkw1132)

23. Waterhouse A, Bertoni M, Bienert S, et al. SWISS-MODEL: homology modelling of protein structures and complexes. *Nucleic Acids Research.* 2018;46(W1): W296-W303. (DOI:10.1093/nar/gky427)

24. Schrodinger L. Small-Molecule Drug Discovery Suite 2019-3. 2017-2 ed. New York, NY: Schrodinger, LLC; 2019.

25. Davis IW, Leaver-Fay A, Chen VB, et al. MolProbity: all-atom contacts and structure validation for proteins and nucleic acids. *Nucleic Acids Res.* 2007;35(Web Server issue): W375-383. (DOI:10.1093/nar/gkm216)

26. The PyMOL Molecular Graphics System, Version 2.0 Schrödinger, LLC.
Supporting Information for

Homology models of Wuhan Coronavirus 3CL\textsuperscript{pro} Protease

Dr. Martin J. Stoermer  
Division of Chemistry and Structural Biology, Institute for Molecular Bioscience  
The University of Queensland  
St. Lucia, 4072, Queensland, Australia  
E-mail: martin.stoermer@uq.edu.au  
Orcid: https://orcid.org/0000-0003-3445-2104  
Twitter: @MartinStoermer

Nomenclature

Original Swissmodel homology models were numbered 1-8. However due to redundancy monomer Models 1,2,4,6,7,8 were taken through to refinement. Model 8 was only obtained as a monomer, so just 5 models were refined as their homodimers. For clarity and to define their point of origin, these were named Models11,12,14,16,17&18.

E.g.:

| Raw Model | PDB Template | Refined monomer model | Refined dimer model |
|-----------|--------------|-----------------------|--------------------|
| 1 (dimer) | 2z9j         | Model 1               | Model 11           |
| 2 (dimer) | 3vb3         | Model 2               | Model 12           |
| 3 (dimer) | 2z9j         | Model 4               | Model 14           |
| 4 (dimer) | 1uk3         | Model 6               | Model 16           |
| 5 (dimer) | 1uk3         | Model 7               | Model 17           |
| 6 (dimer) | 2a5i         | Model 8               |                    |
| 7 (dimer) | 1uj1         |                       |                    |
| 8 (monomer) | 1z1i       |                       |                    |

Table S1. Naming system used for the models at each stage of refinement.
Figure S1. EMBOSS Needle alignment of the extracted protease sequence of Wuhan coronavirus against Bat coronavirus (PDB:2YNB) Catalytic dyads (His41, Cys145) shown in red bold.
Figure S2. EMBOSS Needle alignment of the extracted protease sequence of Wuhan coronavirus against SARS coronavirus (PDB: 2z9j) Catalytic dyads shown in red bold.

MolProbity scoring and statistics for refined monomer models 1,2,4,6,7,8 and dimer models

Monomer models.

| Mode | Template name | Ramachandran favoured (%) | Ramachandran allowed (%) | Clashscore (percentile) | MolProbity score (percentile) | CA Geometry outliers (%) | Cβ deviations (%) |
|------|---------------|--------------------------|--------------------------|-------------------------|-------------------------------|-------------------------|------------------|
| 1    | 2z9j.1.B      | 96.33                    | 100                      | 1.52 (99<sup>a</sup>)   | 1.36 (98<sup>a</sup>)        | 0.0                     | 0.0              |
| 2    | 3vb3.1.A      | 96.66                    | 100                      | 1.30 (99<sup>a</sup>)   | 1.12 (100<sup>a</sup>)       | 0.0                     | 0.0              |
| 4    | 1uk3.1.B      | 93.98                    | 99.67                   | 1.52 (99<sup>a</sup>)   | 1.74 (88<sup>a</sup>)         | 0                      | 0               |
| Model | Template name | Ramachandran favoured (%) | Ramachandran allowed (%) | Clashscore (percentile) | MolProbity score (percentile) | CA Geometry outliers (%) | Cβ deviations (%) |
|-------|---------------|---------------------------|--------------------------|-------------------------|------------------------------|--------------------------|-------------------|
| 6     | 2a5i.1.A      | 96.38                     | 99.67                    | 0.63 (99th)             | 0.95 (100th)                | 0                        | 0                 |
| 7     | 1uj1.1.B      | 96.66                     | 99.67                    | 1.52 (99th)             | 1.43 (96th)                 | 0                        | 0                 |
| 8     | 1zl1.1.A      | 91.30                     | 99.33                    | 3.04 (98th)             | 2.02 (75th)                 | 0.34                     | 0.36              |

Table S2. Statistics via MolProbity with no His/Asn/Gln flips allowed.

| Model | Template name | Ramachandran favoured (%) | Ramachandran allowed (%) | Clashscore (percentile) | MolProbity score (percentile) | CA Geometry outliers (%) | Cβ deviations (%) |
|-------|---------------|---------------------------|--------------------------|-------------------------|------------------------------|--------------------------|-------------------|
| 1 (1 flip) | 2z9j.1.B | 96.33                     | 100                      | 1.30 (99th)             | 1.32 (98th)                 | 0                        | 0                 |
| 2 (2 flips) | 3vb3.1.A | 96.66                     | 100                      | 1.30 (99th)             | 1.12 (100th)                | 0                        | 0                 |
| 4 (0 flips) | 1uk3.1.B | 93.98                     | 99.67                    | 1.52 (99th)             | 1.74 (88th)                 | 0                        | 0                 |
| 6 (2 flips) | 2a5i.1.A | 96.38                     | 99.67                    | 0.63 (99th)             | 0.95 (100th)                | 0                        | 0                 |
| 7 (1 flip) | 1uj1.1.B | 96.66                     | 99.67                    | 1.52 (99th)             | 1.43 (96th)                 | 0                        | 0                 |
| 8 (3 flips) | 1zl1.1.A | 91.30                     | 99.33                    | 3.04 (98th)             | 2.02 (75th)                 | 0.34                     | 0.36              |

Table S3. Statistics via MolProbity with His/Asn/Gln flips allowed. 5/6 models had the possibility of suggested hydrogen bonding flips, but only the top model1 resulted in any improvement to the clash and MolProbity scores.

Dimer models:

| Dimer | Template name | Ramachandran favoured (%) | Ramachandran allowed (%) | Clashscore (percentile) | MolProbity score (percentile) | CA Geometry outliers (%) | Cβ deviations (%) |
|-------|---------------|---------------------------|--------------------------|-------------------------|------------------------------|--------------------------|-------------------|
| 11    | 2z9j.1.B      | 96.51                     | 99.83                    | 1.62 (99th)             | 1.41 (97th)                 | 0.17                     | 0                 |
| 12    | 3vb3.1.A      | 96.35                     | 100                      | 1.18 (99th)             | 1.12 (100th)                | 0.0                      | 0                 |
| 14    | 1uk3.1.B      | 94.16                     | 99.67                    | 1.84 (99th)             | 1.73 (88th)                 | 0.34                     | 0                 |
| 16    | 2a5i.1.A      | 95.72                     | 99.67                    | 0.73 (99th)             | 1.03 (100th)                | 0                        | 0                 |
| 17    | 1uj1.1.B      | 94.82                     | 99.83                    | 1.30 (99th)             | 1.58 (93th)                 | 0                        | 0.18              |

Table S4. Statistics via MolProbity with no His/Asn/Gln flips allowed.
Table S5. Statistics via MolProbity with His/Asn/Gln flips allowed.
96.3% (289/300) of all residues were in favored (98%) regions.
100.0% (300/300) of all residues were in allowed (>99.8%) regions.

**Figure S2.** Ramachandran plots for Wuhan CoV 3CL\textsuperscript{pro} homology monomer model1
96.7% (289/299) of all residues were in favored (98%) regions.
100.0% (299/299) of all residues were in allowed (>99.8%) regions.

**Figure S3.** Ramachandran plots for Wuhan CoV 3CL<sup>pro</sup> homology monomer model 2
MolProbity Ramachandran analysis

WH_human_protease_from1uk3_Minimised1H.pdb, model 4

94.0% (281/299) of all residues were in favored (98%) regions.
99.7% (298/299) of all residues were in allowed (>99.8%) regions.

Figure S4. Ramachandran plots for Wuhan CoV 3CL<sup>pro</sup> homology monomer model4
MolProbity Ramachandran analysis

WH_human_protease_from2a5i_Minimised1H.pdb, model 6

96.4% (293/304) of all residues were in favored (98%) regions.
99.7% (303/304) of all residues were in allowed (>99.8%) regions.

Figure S5. Ramachandran plots for Wuhan CoV 3CLpro homology monomer model 6
MolProbity Ramachandran analysis

WH_human_protease_fromluj1_Minimised1H.pdb, model 7

96.7% (289/299) of all residues were in favored (98%) regions.
99.7% (298/299) of all residues were in allowed (>99.8%) regions.

Figure S6. Ramachandran plots for Wuhan CoV 3CL\textsuperscript{pro} homology monomer model 7
MolProbity Ramachandran analysis

WH_human_protease_from1z1i_Minimised1H.pdb, model 8

91.3% (273/299) of all residues were in favored (98%) regions.
99.3% (297/299) of all residues were in allowed (>99.8%) regions.

Figure S7. Ramachandran plots for Wuhan CoV 3CL\textsuperscript{pro} homology monomer model8
96.5% (581/602) of all residues were in favored (98%) regions. 99.8% (601/602) of all residues were in allowed (>99.8%) regions.

Figure S8. Ramachandran plots for Wuhan CoV 3CLpro homology dimer model11
MolProbity Ramachandran analysis

**WH_human_protease_from3vb3_dimerMinimised1H.pdb, model 12**

96.4% (581/603) of all residues were in favored (98%) regions.
100.0% (603/603) of all residues were in allowed (>99.8%) regions.

**Figure S9.** Ramachandran plots for Wuhan CoV 3CL\textsuperscript{pro} homology dimer model12
MolProbity Ramachandran analysis

WH_human_protease_from_luk3_dimerMinimised1H.pdb, model 14

94.2% (564/599) of all residues were in favored (98%) regions.
99.7% (597/599) of all residues were in allowed (>99.8%) regions.

Figure S10. Ramachandran plots for Wuhan CoV 3CL\textsuperscript{pro} homology dimer model14
MolProbity Ramachandran analysis

WH_human_protease_from2a5i_dimerMinimised1H.pdb, model 16

95.7% (582/608) of all residues were in favored (98%) regions.
99.7% (606/608) of all residues were in allowed (>99.8%) regions.

Figure S11. Ramachandran plots for Wuhan CoV 3CL\textsuperscript{pro} homology dimer model16
94.8% (568/599) of all residues were in favored (98%) regions.
99.8% (598/599) of all residues were in allowed (>99.8%) regions.

Figure S12. Ramachandran plots for Wuhan CoV 3CL\textsuperscript{pro} homology dimer model17
Molecular Dynamics Method

Molecular dynamics simulations were performed with the Desmond Molecular Dynamics System (D. E. Shaw Research, New York, NY) by using the tools incorporated in the Schrödinger Suite 2019-2. The model system was set up by using the “System Setup” utility. The OPLS2005 force field parameters were implemented during simulations. Each monomer, dimer or ternary complex was placed in an orthorhombic box with solvent buffers up to 10Å along each dimension. The TIP3P solvent model was used to describe water molecules. Overall neutralisation of the system was achieved by adding sodium and chloride ions at the physiological concentration of 0.15M. The prepared model systems were then relaxed by using the multi-step default protocol in Desmond. After relaxation, the whole system was subjected to a 100ns simulation. The cut-off distance for computing short-range electrostatics and Lennard–Jones interaction was set to 9.0Å. The trajectories and energies were recorded every 100 and 1.2ps, respectively. MD simulations were performed at the IMB cluster from the University of Queensland on NVIDIA P100 GPUs or on a CentOS Linux Desktop with a Quadro RTX 5000 GPU. Simulation analysis was performed on Linux or Macintosh desktops using the Simulation Interactions Diagrams module in Schrödinger Suite 2019-2.

Model2 from SARS 3vb3 WH_human_protease_from3vb3_Minimised.pdb

Figure S1-13. Protein C-alpha RMSD over the 100 ns duration molecular dynamics simulation. Figure S1-14. Protein C-alpha RMSF plot. Per residue variability in Cα position.

Flexible residues

Figure S1-15. WH CoV 3CLPRO Model 2 shown as green ribbons, except catalytic dyad (His41, Cys145 – yellow sticks), and magenta ribbons for regions found to be mobile in 100 ns MD simulation.
Moderately mobile regions (>1.5 RMSF) N-terminal residues Ser1-Phe3, β-hairpin loop Gly23-Thr25, α-helix and loop Thr45-Arg60, loop tip Asn72, strand Gly138-Gly143, β-hairpin loop Pro168-Gly170, long strand Glu189-Gly195, strand Asn274-Gly278, C-terminus Gln299-Ser301.

Model6 from SARS 2a5i WH_human_protease_from2a5i_Minimised.pdb

Figure S1-16. Protein C-alpha (blue line) and ligand (red line) RMSD over the 100 ns duration MD run. Figure S1-17 Protein RMSF plot. Per residue variability in Cα position. Figure S1-18. Protein- ligand interaction summary diagram. Orange: negative charge interaction, blue: polar interactions, green: hydrophobic interactions, grey circles: solvent exposed atoms.

Over the first 10 ns of the simulation the mobile loop 44CTSEDMLNPN53 defining the boundary of the S2 subsite moves away from the bound ligand, and remains distant during the remaining 80ns.
Flexible residues

**Figure S1-19.** WH CoV 3CL\textsuperscript{pro} Model 6 shown as green ribbons, except catalytic dyad (His41, Cys145 – yellow sticks), and magenta ribbons for regions found to be mobile in the 100 ns molecular dynamics simulation. Covalently bound SARS inhibitor carried through in model from template PDB:2a5i shown as grey sticks.

Moderately mobile regions (>1.5 RMSD) N-terminal residues Ser1-Gly2, β-hairpin loop Gly23-Thr25, very mobile α-helix and loop Thr45-Arg60, loop tip Asn72, long strand Arg188-Ala194, strand Arg222-Thr224, strand Asn274-Arg279, C-terminus Ser301-Gln306.

**Model12 from SARS 3vb3 WH_human_protease_from3vb3_dimerMinimised.pdb**

**Figure S1-20.** Protein C-alpha (blue line) RMSD over the 100 ns duration MD run. **Figure S1-21.** Protein RMSF plot. Per residue variability in C\(\alpha\) position.
Flexible residues

**Figure S1-21.** WH CoV 3CL\textsuperscript{pro} Model 12 Chain A shown as green ribbons, Chain B cyan ribbons, except catalytic dyads (His41, Cys145 – green and cyan sticks), and magenta ribbons for regions found to be mobile in 100 ns MD simulation.

In comparison to the MD simulation of the monomer (Model 2), the dimer shows stable N- and C-termini for both subunits, as these form part of the dimer interface which remains quite stable during the simulation. The very mobile helix and loop region show slight differences in mobility (Chain A Ser46-Ser62, Chain B Thr45-Asp56). In addition there are differences in the N-terminal domains, the Chain B helices Leu227-Met235 and Glu270-Met276 being slightly more mobile than those in Chain A.

Moderately mobile regions (>1.5 RMSD) Chain A: \(\alpha\)-helix and loop Ser46-Ser62, loop tip Asn72, long strand 215-Thr225, helical turn Thr243-His246, large loop Asn274-Arg279, C-terminus Ser301. Chain B: \(\beta\)-hairpin loop Gly23-Thr25, \(\alpha\)-helix and loop Thr45-Asp56, loop tip Asn72, loop Pro168-Gly170, long strand and helix Trp218-Met235, loop Thr243-Gln244, loop and helical turn Glu270-Thr280, loop Gly283-Leu286.

Model16 from SARS 2a5i WH\_human\_protease\_from2a5i\_dimerMinimised.pdb

**Figure S1-22.** Protein C-alpha (blue line) and ligand (red line) RMSD over the 100 ns duration MD run. **Figure S1-23.** Protein RMSF plot. Per residue variability in C\(\alpha\) position.
**Figure S1-24.** Protein-ligand interaction summary diagram. Orange: negative charge interaction, blue: polar interactions, green: hydrophobic interactions, grey circles: solvent exposed atoms.

**Flexible residues**

**Figure S1-14.** WH CoV 3CL\textsuperscript{pro} Model 16 Chain A shown as green ribbons, Chain B cyan ribbons, except catalytic dyads (His41, Cys145 – yellow sticks), and magenta ribbons for regions found to be mobile in 100 ns MD simulation. SARS 3CL\textsuperscript{pro} inhibitor shown as grey sticks.
Apart from the more stable N- and C-termini for both subunits which form part of the dimer interface the remainder of the dimer remains of similar overall flexibility. One major difference is that in the dimeric form the loop $^{44}$CTSEDMLNP$^{53}$ which forms the boundary of the S2-subsite which binds bulky substrate residues Met/Phe/Leu is much less mobile than was seen in the 100ns simulation of the monomer.

Moderately mobile regions (>1.5 RMSD) Chain A: loop Thr45-Asp48, $\alpha$-helix Asn53-Arg60, loop tip Asn72, loop tip Tyr154, strand Ala191-Ala193, strand residue Phe223, strand Asn274-Gly278, C-terminus Gly302-Gln306. Chain B: $\beta$-hairpin loop Gly23-Thr24, large loop Cys44-Asp53, strand Ile59-His64, loop tip Asn72, loop tip Tyr154, long strand and loop Val186-Gly195, strand residue Gly215, strand Phe223-Thr224, $\alpha$-helical turn Leu232-Tyr237, helix residue Val247, helical turn Ala255-Gly258, very mobile loop and helical turn Glu270-Leu286, C-terminus Gly302-Gln306
| File Name                                      | Size          | Actions                      |
|-----------------------------------------------|---------------|------------------------------|
| EMBOSS_NEEDLE_Alignments.tgz                   | 3.26 KiB      | view on ChemRxiv, download file |
| Raw_models.tgz                                | 1.19 MiB      | view on ChemRxiv, download file |
| Refined_Dimers.tgz                            | 813.48 KiB    | view on ChemRxiv, download file |
| Refined_monomers.tgz                          | 489.68 KiB    | view on ChemRxiv, download file |
| Bat_WH_SARS_protease.clustal                   | 1.77 KiB      | view on ChemRxiv, download file |
| README.txt                                    | 1.21 KiB      | view on ChemRxiv, download file |
| WH_human_3CLproModel2_from3vb3_Minimised.pdb   | 367.99 KiB    | view on ChemRxiv, download file |