



# Evolutionary aspects of mutation in functional motif and post-translational modifications in SARS-CoV-2 3CLpro (Mpro): an in-silico study

Amit Gupta<sup>1</sup> · Niharika Sahu<sup>1</sup> · Vinay Kumar Singh<sup>2</sup> · Rajeshwar P. Sinha<sup>1,3</sup>

Received: 20 December 2022 / Revised: 20 March 2023 / Accepted: 30 March 2023  
© The Author(s), under exclusive licence to Springer Nature Singapore Pte Ltd. 2023

## Abstract

SARS CoV-2 is the virus that caused the COVID-19 pandemic. The main protease is one of the most prominent pharmacological targets for developing anti-COVID-19 therapeutic drugs (Mpro); SARS-CoV-2 replication is dependent on this component. SARS CoV-2's Mpro/cysteine protease is quite identical to SARS CoV-1's Mpro/cysteine protease. However, there is limited information on its structural and conformational properties. The present study aims to perform a complete in silico evaluation of Mpro protein's physicochemical properties. The motif prediction, post-translational modifications, effect of point mutation, and phylogenetic links were studied with other homologs to understand the molecular and evolutionary mechanisms of these proteins. The Mpro protein sequence was obtained in FASTA format from the RCSB Protein Data Bank. The structure of this protein was further characterized and analyzed using standard bioinformatics methods. According to Mpro's in-silico characterization, the protein is a basic, non-polar, and thermally stable globular protein. The outcomes of the phylogenetic and synteny study showed that the protein's functional domain amino acid sequence is substantially conserved. Furthermore, it has undergone many changes at the motif level over time from porcine epidemic diarrhoea virus to SARS-CoV 2, possibly to achieve various functions. Several post-translational modifications (PTMs) were also observed, and the possibilities of changes in Mpro protein exhibit additional orders of peptidase function regulation. During heatmap development, the effect of a point mutation on the Mpro protein was seen. This protein's structural characterization will aid in a better understanding of its function and mechanism of action.

**Keywords** SARS-CoV-2 · Mpro · In silico analysis · Heatmap · Phylogenetic relationship · Post-translational modifications (PTMs)

## Introduction

The global transmission of a novel coronavirus strain (Chakraborty et al. 2020) began in Wuhan, China, at the end of 2019 and rapidly acquired pandemic proportions. The first line of defence is global vaccination programs and efforts to

produce oral medication therapies (Rothan and Byrareddy 2020). The virus-caused pandemic disease was termed "Coronavirus disease 2019" by the World Health Organization in the spring 2021 (<https://covid19.who.int/>). The novel virus was given the name "severe acute respiratory syndrome coronavirus-2" by the International Committee on Virus Taxonomy (<https://talk.ictvonline.org/>). COVID-19 infection by inhalation has been linked to a variety of respiratory symptoms, ranging from minor symptoms such as coughing, headaches, and fever to pneumonia and severe breathing impairment, sometimes leading to fulminant death in certain types of patients, even when treated in hospital intensive care units (ICUs) (Zhao et al. 2020). SARS CoV-2, a member of the "Coronaviridae family and genus b-coronavirus, is a member of the Coronaviridae family and genus b-coronavirus" (Zhou et al. 2020).

✉ Rajeshwar P. Sinha  
rpsinhabhu@gmail.com

<sup>1</sup> Laboratory of Photobiology and Molecular Microbiology, Centre of Advanced Study in Botany, Institute of Science, Banaras Hindu University, Varanasi 221005, India

<sup>2</sup> Centre for Bioinformatics, School of Biotechnology, Institute of Science, Banaras Hindu University, Varanasi 221005, India

<sup>3</sup> University Center for Research and Development (UCRD), Chandigarh University, Chandigarh, India

Since 2003, three coronaviruses have been linked to pneumonia. Between 2002 and 2003, SARS CoV-1 infected 8098 people and killed 774 of them (Drosten et al. 2003; Walls et al. 2020). The Middle East respiratory syndrome coronavirus was the second (MERS-CoV), affecting 27 nations and infecting 2494 people. MERS CoV caused 858 deaths (Zaki et al. 2012). SARS CoV-1 and SARS CoV-2 were revealed to be genetically identical, sharing almost 80% of their genomes (Zhou et al. 2020). Furthermore, both coronaviruses are assumed to have come from the same source (bats), implying the same reservoir host (Zhou et al. 2020). SARS CoV-1 and other coronaviruses, such as MERS CoV, have provided insight into COVID-19's virology. SARS CoV-2, like the other two coronaviruses, is an enclosed virus with a large 30 kb single-stranded positive-sense RNA genome. The viral gene essential for virus replication is a polyprotein that should be broken up into functional components in order for replication followed by transcription to occur. (Zhang and Wang 2021). Cleavage occurs mostly after a Gln residue, and a hydrophobic residue usually follows the Gln residue, most commonly Leu. The residue after the Gln is frequently a minor amino acid like Ser, Ala, or Asn. Mpro detaches itself from the polyprotein by autolysis (Hegyi and Ziebuhr 2002). Limiting or inhibiting viral replication by inhibiting Mpro activity can enhance clinical significance for coronavirus and other disorders. However, no such human proteases with similar cleavage selectivity to Mpro have been identified, implying that developing inhibitors that target Mpro without causing off-target damage should be conceivable (Kim et al. 2016).

The main protease (Mpro) of SARS-CoV-2 has been the target of many therapeutic candidates (Kneller et al. 2020; Joshi et al. 2021; Zhang et al. 2020). Because these proteins are synthesized in the form of polyproteins that must be released to produce mature proteins, this enzyme plays a critical role in virus propagation (Kneller et al. 2020). The transcription as well as the replication of viral RNAs is both carried out by this multifunctional protein like polyprotein 1a, and Mpro (nsp5) is a monomeric part of it (Fig. S1). The proteinases that cause the polyprotein to be cleaved are found here. It is responsible for a large portion of the proteolytic cleavage required to obtain the functional proteins that are essential for virus replication and transcription. SARS-CoV In solution, Mpro is made up of combinations of monomers and dimers; its active state has been linked to its dimeric form (Joshi et al. 2021). The major protease (Mpro) sequence is substantially conserved across coronavirus species (Mirza and Froeyen 2020). According to studies, the Mpro from SARS CoV-2 has a sequence similarity of more than 96% to the same protease identified in MERS and SARS CoV-1, as a result, it is a good candidate for broad-spectrum anti-CoV treatment (Ghosh et al. 2021).

The Mpro of SARS CoV-1 and SARS CoV-2 have much more sequence similarities.

The majority of current research has now focused on reducing the enzymatic activity of SARS CoV-2 Mpro's major target protein by explicitly targeting its active site with target inhibitory drugs that compete for active site "pockets" (Zhao et al. 2020; Zhang et al. 2020; Dai et al. 2020). Studies have been conducted using a drug repurposing technique to discover appropriate inhibitors of Mpro (Bharadwaj et al. 2020; Jiménez-Alberto et al. 2020). Despite information on the Mpro protein's activity in proteases, its role beyond the proteases and the molecular mechanisms involved by Mpro protein is unknown. As a result, a thorough understanding of the Mpro protein's structure and confirmation is critical. Given the significance of this protein, this work aims to examine its structural characteristics, including physicochemical properties, tertiary structure analysis, motif prediction, post-translational modifications, Heatmap generation and evolutionary relationships.

## Materials and methods

### Sequence and structure retrieval and Mpro protein preparation

The SARS CoV-2 Mpro crystal structure and the protein sequence were retrieved in the FASTA format from the RCSB Protein Data Bank (<http://www.rcsb.org>). Protein preparation was done with the help of the 'Prepare protein' protocol of BIOVIA Discovery studio. Heteroatoms and water molecules present in the crystal structure were removed at physiological pH 7.4. Further, the active site prediction of the prepared protein was made by CASTp 3.0 server (<http://sts.bioe.uic.edu/>) (Tian et al. 2018).

### Physicochemical characterization of selected Mpro protein

The grand average of hydropathicity value (GRAVY), molecular weight, instability index, aliphatic index, isoelectric point (pI), and molecular weight of selected target protein were calculated using the ExPASy ProtParam server (<https://web.expasy.org/protparam/>) (Gasteiger et al. 2005).

### Sequence comparison and phylogenetic analysis of Mpro protein

The PSI-BLAST (<https://blast.ncbi.nlm.nih.gov/Blast.cgi?PAGE=Proteins>) tool was used to search the protein homologs available in the NCBI data model. Jalview's MUSCLE with defaults algorithm was used to align all of the sequences. (<http://www.jalview.org/>). The aligned

sequences were color-coded as retained hydrophobicity, and a ten percent cut-off was used (Altschul et al. 1990, 1997; Waterhouse et al. 2009). The sequences of the Mpro protein homologs available in the NCBI database were retrieved in the FASTA format; The ClustalW method with the default algorithm was then used to align all of the sequences. The alignment session was then saved and used to build the phylogenetic tree (Neighbor-Joining Tree) with a bootstrap value of 1000 pseudo replicas (Sinha and Meller 2007, 2008; Kumar et al. 2018). The MEGA X software was used to construct the phylogenetic tree. The Seq2Logo server was used for the visualization of aligned sequences (<http://www.cbs.dtu.dk/biotools/Seq2Logo/>).

### Domain analysis and motif identification

The Mpro protein functional domain was predicted by the InterproScan (<http://www.ebi.ac.uk/Tools/pfa/iprscan>) server. The Prosite (<https://prosite.expasy.org>) server was used for protein domain and functional site analysis (Sigrist et al. 2012). The identified Mpro proteins homologs were subjected to MEME tool search (<http://meme-suite.org/tools/meme>), for the prediction of conserved motifs of Mpro proteins (Bailey et al. 2015), Motif search tool (<http://www.genome.jp/tools/motif>) was then searched for the functional characteristics of identified common motifs in the database of known proteins.

### Post-translational modifications (PTMs) prediction, antigenicity, and heatmap generation of point mutation

Post-translational Modifications like phosphorylation, glycosylation, methylation and sumoylation Mpro were predicted by NetPhos v.3.1 (<http://www.cbs.dtu.dk/services/NetPhos>) NetNGlyc1.0Server was used for glycosylation prediction (<http://www.cbs.dtu.dk/services/NetNGlyc>), for methylation prediction PRmePRed was used (<http://bioinfo.icgeb.res.in/PRmePRed/>) SUMOPlotTM server was used for sumoylation prediction (<https://www.abcepta.com/sumoplot>) (Xue et al. 2006) servers, respectively. Protective antigenicity prediction was made Vaxijen v2.0 online server (<http://www.ddgpharmfac.net/Vaxijen/Vaxijen/VaxiJen.html>). Effect of point mutation was predicted using Predict Protein tool (<https://predictprotein.org/>) (Yachdav et al. 2014).

### Tertiary structure analysis, validation, the secondary structure arrangement, and topology

Among the 29 sequences (PDB ID: 6LU7) was selected as a representative of all to predict the tertiary structure of Mpro protein. The 3D models of the Mpro protein sequence (PDB ID: 6LU7) with energy minimization parameters were

selected (Arnold et al. 2006). PyMOL (Schrödinger Inc.) was used to view the model's publishable picture (DeLano 2002; Gupta et al. 2022). The most crucial aspect of structure prediction was structure evaluation. ProSA (<https://prosa.services.came.sbg.ac.at/prosa.php>) (Sippl 1993) and VADAR (<http://vadar.wishartlab.com>) (Willard et al. 2003) web servers were used to assess the tertiary protein structures. The validation of the model was done by using PROCHECK (<https://servicesn.mbi.ucla.edu/PROCHECK/>), secondary structure, Ramachandran plot, and topology of protein were predicted by PDBsum online tool (<http://www.ebi.ac.uk/thornton-srv/databases/pdbsum/Generate.html>).

## Results and discussion

### Sequence retrieval and preparation of Mpro

SARS CoV-2 Mpro is one of the proteinases that cleave the polyprotein. Replicase polyprotein 1ab is a multifunctional protein that is involved in viral RNA transcription and replication (Yoshimoto 2020). Mpro protein being major proteinases, which cleaves the replicase polyprotein's C-terminus at 11 different sites. There was a strong reason behind its selection as a candidate protein. The protein sequence was retrieved in the FASTA format using RCSB Protein Data Bank (PDB ID: 6LU7) (Jin et al. 2020). It has only one chain of a total of 306 amino acids. The protein was prepared using the BIOVIA Discovery studio's 'Prepare protein' methodology. At physiological pH 7.4, heteroatoms water molecules were removed from the crystal structure. In addition, the active site of the prepared protein was predicted. Amino acid residues of THR<sup>25</sup>, HIS<sup>41</sup>, MET<sup>49</sup>, PHE<sup>140</sup>, ASN<sup>142</sup>, GLY<sup>143</sup>, SER<sup>144</sup>, MET<sup>165</sup>, GLU<sup>166</sup>, and GLN<sup>189</sup> were found to be the active site residues of Mpro protein (Fig. S2).

### Primary structural analysis of Mpro

The ExPASy ProtParam server was used to assess the amino acid content (Table 1) and physicochemical characteristics (Table 2) of the Mpro protein. The Mpro protein consists of 306 amino acid (A.A.) residues. Protein's theoretical isoelectric point (pI) was calculated to be 5.95, which shows that the Mpro protein is acidic. Because there were more negatively charged A.A. residues than positively charged A.A. residues, it was determined to be acidic. The Mpro protein was observed to have an instability score of less than 40, indicating that it is a highly stable protein (Singh et al. 2021). The aliphatic index of the Mpro protein was computed to be 82. This could be viewed as a beneficial element in boosting Mpro protein thermostability, the relative volume filled by a protein's aliphatic side chains is primarily characterized by its aliphatic index (alanine, valine, leucine,

**Table 1** Amino acid composition of the SARS-COV 3CL Pro protein

S. no.	A.A.	Quantity	Percentage (%)
1.	Ala (A)	17	5.6
2.	Arg (R)	11	3.6
3.	Asn (N)	21	6.9
4.	Asp (D)	17	5.6
5.	Cys (C)	12	3.9
6.	Gln (Q)	14	4.6
7.	Glu (E)	9	2.9
8.	Gly (G)	26	8.5
9.	His (H)	7	2.3
10.	Ile (I)	11	3.6
11.	Leu (L)	29	9.5
12.	Lys (K)	11	3.6
13.	Met (M)	10	3.3
14.	Phe (F)	17	5.6
15.	Pro (P)	13	4.2
16.	Ser (S)	16	5.2
17.	Thr (T)	24	7.8
18.	Trp (W)	3	1.0
19.	Tyr (Y)	11	3.6
20.	Val (V)	27	8.8

**Table 2** The Physico-chemical characteristics of the SARS-COV 3CLPro protein

S. no.	Physicochemical characteristics	Values
1.	Number of amino acids	306
2.	Molecular weight	33,796.64
3.	Theoretical pI	5.95
4.	Number of negatively charged residues (Asp + Glu) total	26
5.	Number of positively charged residues (Arg + Lys) total	22
6.	Total number of atoms	4686
7.	Extinction coefficient	32,890
8.	Instability index	27.65
9.	Aliphatic index	82.12
10.	Grand average of hydropathicity (GRAVY)	−0.019

and isoleucine). In addition, this protein's GRAVY value was reported to be negative, suggesting that it is non-polar.

## Sequence and phylogenetic analysis

The coronavirus main protease (M-pro) domain profile from SARS-CoV-2 R1A\_SARS2 Replicase polyprotein 1a (UniProt ID: P0DTC1) was taken for sequence and phylogenetic analysis. The selected candidate Mpro protein sequence alignment was done with homologs 28 sequences. The sequence similarity was estimated (lowest score of 228

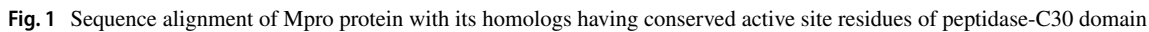
and *e*-value of  $3e-73$ ) with Mpro protein of SARS-CoV 2 revealed that the Peptidase-C30 domain was highly conserved throughout homologs sequences (Fig. 1). Multiple sequence alignment (MSA) of 29 sequences revealed that the protein contains conserved amino acid residues (Fig. S3). While ten proline residues out of thirteen found were not conserved in Mpro, three were highly conserved throughout sequences. Proline residue conservation is found to be less, which suggests that preserved structural rigidity is less in terms of coiled-coil structure, and it has its own functional importance (Yu et al. 2015). The proline's side chain's distinct cyclic shape likely contributed conformational stiffness, but in Mpro protein, no conformational rigidity was found because of lesser proline content. Furthermore, the peptide bond's Phi/Psi angles showed allowed degrees of rotation, resulting in unconstrained conformational space for proline residues in the main chain (Xue et al. 2006).

Further, the proline-assisted folding in sites where the main chain conformation was found to be compatible with the permissible range. Figure 1 shows the sequence alignment of the Mpro protein with the homologs with highlighted conserved amino acids. The creation of various clusters occurred due to evolutionary changes in amino acid sequences depicted in a phylogenetic tree. To create the phylogenetic tree, amino acid sequences were subjected to the Neighbor-Joining Tree (MEGA X) with 1000 bootstrapping resampling. The tested Mpro protein generated two main clusters in the phylogenetic tree (Fig. 2). Mpro developed into two primary groups, according to a dendrogram created using the protein. Later on, these two main clusters formed in a phylogenetic tree are explained based on motif differentiation. The evolution of the Mpro protein from the porcine epidemic diarrhea virus to SARS-CoV2 may be due to some deletion/insertion in the Mpro protein-coding gene. In due course of evolution in Mpro protein, it was observed that MERS-CoV and Bat-CoV showed a similar motif pattern differentiating themselves from others. The positional conservation of Mpro protein was identical in Bat-CoV and SARS-CoV-2 (humans), indicating that the gene has gone through a transitional period during genome organization that occurred with the evolution. The most evolved Mpro protein was observed in SARS-CoV and SARS-CoV-2. This protein was remarkably similar in diverse organisms, demonstrating its functional conservation throughout the course of evolution.

## Domain analysis and motif identification

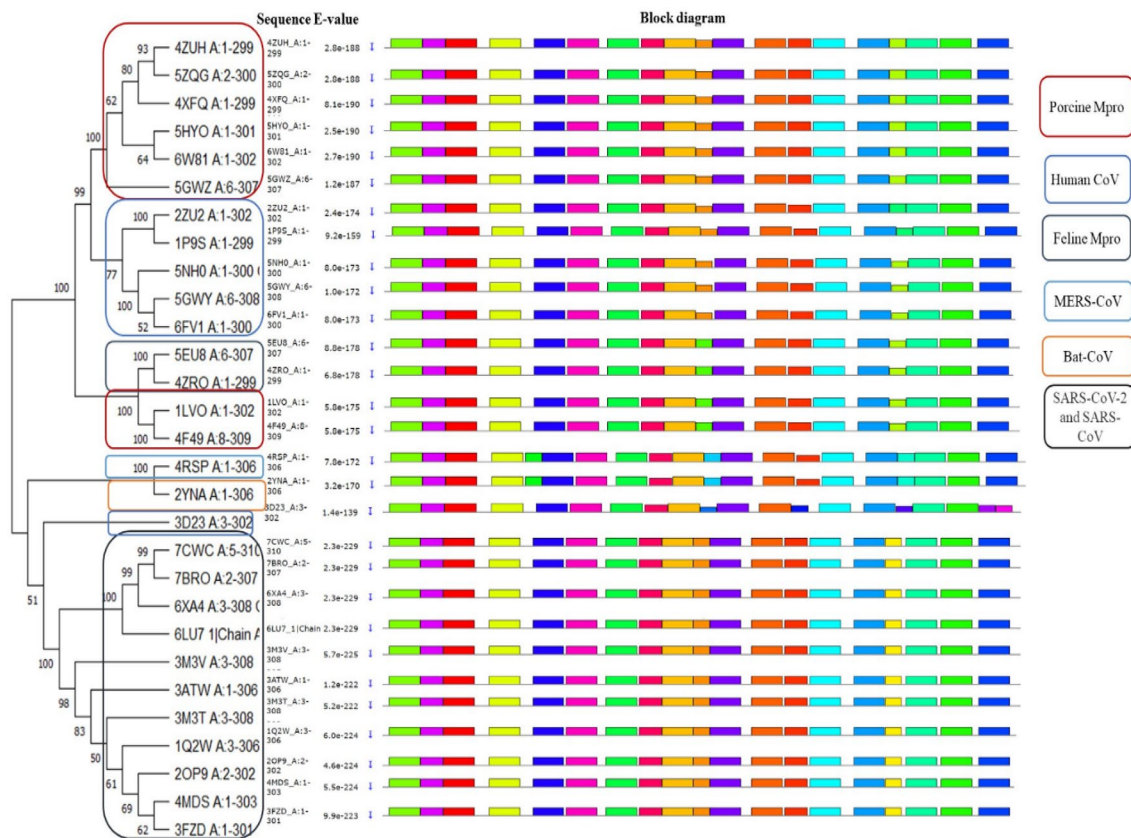
The primary structure analysis of the M-pro protein included motif and domain analysis. Viral RNA transcription and replication are both mediated by multifunctional proteins, which has the proteinases needed for cleavages of the poly-protein. Peptidase C30 (1–306) amino acid, SARS-CoV-2





broad antiviral activity. A total of 30 motifs were observed in 29 sequences when subjected to MEME and depicted in Fig. 2 shows the motifs with their width and the best potential amino acid sequence matches. It was observed that 15 motifs out of 30 are fully conserved, and the majority of them revealed their identity as the Peptidase-C30 domains (Fig. 2 and Fig. S5). Mainly larger size motifs are conserved, whereas comparatively smaller size motifs were not conserved. A set of 15 amino acid residues in motif 1 (30–44), motif 4 (158–172), and motif 5 (135–149) amino acids positions, i.e., **LWLGD TVYCPRH VIA**, **EFCYMH HMELGTGVH**, and **TIKGSFJNGSCGSVG** were uniformly observed, and these motifs were associated with the active site of the protein. Motif differentiations were found to be a major basis of cluster formation in the phylogenetic tree.

The phylogenetic tree revealed the formation of two major branches, branches 1 and 2 that have several sub-branches. Whereas motifs are the basis for the explanation of the phylogenetic tree, some motifs were selected for it. Motif 21 was observed in (PDB ID: 5EU8, 4ZRO, 1LVO, and



**Fig. 2** Phylogenetic tree of selected 29 different Mpro proteins of viral origin using Neighbor-joining method of MEGA X along with identified 30 tandem motifs result elucidated by MEME analysis

4F49) belonging to the Peptidase\_C30 domain. It showed a higher amino acid percentage similarity between the Feline infectious peritonitis virus and Porcine transmissible gastroenteritis coronavirus. The motif 25 (YTEEGGVI) was observed in (PDB ID: 4RSP and 2YNA), which depicts the close similarity between MERS-CoV and Bat-CoV Mpro, furthermore motif 29 (GCAFEETEE) and 30 (GPADLEDZ) was observed in (PDB ID: 3D23), which was mainly found in Human coronavirus HKU1. SARS-CoV-2 and SARS-CoV motif 19 (LTQDHVDI) were commonly seen, which showed that both of them have higher amino acid similarity and are closely related.

### Post-translational modifications (PTMs) and heatmap generation of point mutation

Viruses have a broad spectrum of PTMs, which demonstrate differences in protein functions, primarily by influencing cellular processes by affecting 3D conformation and substrate interactions (Szymanski et al. 2002). As a result, protein PTMs have called into question the traditional "one gene, one product" model, which implied that the protein had a single function (Cain et al. 2014; Kharwar et al. 2021).

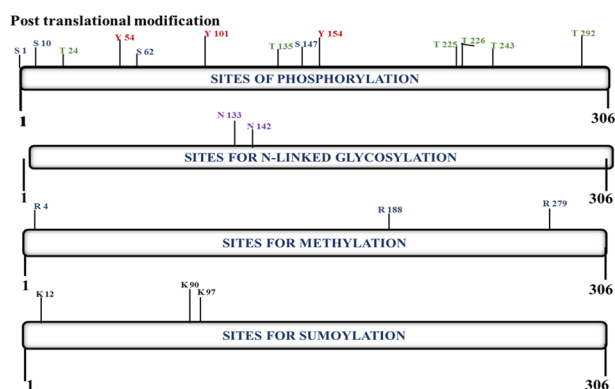
PTM exploratory research is helpful in gaining a holistic and thorough description of the protein, understanding the signal transduction network, and revealing the protein's biological activities in viruses.

Phosphorylation, N-linked glycosylation, methylation, and sumoylation sites, as well as the presence of ubiquitination sites, were predicted using in silico analysis of PTMs in Mpro protein. Phosphorylation occurred extensively in the 3 serine (S: 1, 10, and 62) and 6 threonine's (T: 24, 135, 225, 226, 243, and 292) residues, respectively. In addition, N-linked glycosylation, methylation, and sumoylation occurred in 2 asparagine (N: 133 and 142), 3 arginine (R: 4, 188, and 279), and 3 lysine's (K: 12, 90 and 97) residues, respectively (Fig. 3). Phosphorylation of proteins was discovered to be dynamic in microorganisms, responding to various environmental conditions, including food availability, light, and osmolarity, demonstrating its significance in regulating various important functions such as stress adaptation and C and N metabolism (Zhang et al. 2005).

It's possible that SARS-CoV-2 enzymes (such as Nsp3 and Nsp5) use host proteins as substrates to influence the host cell's signalling system for their own gain by enhancing the diversity and activities of viral proteins

**Table 3** The coronavirus main protease (M-pro) domain profile from SARS-CoV-2 R1A\_SARS2 Replicase polyprotein 1a (UniProt ID: P0DTC1)

S. no.	ScanProsite ID	Name	Start position	End position	Details
1.	PS51962	COV_NSP1	12	133	Coronavirus (CoV) Nsp1 globular domain profile
2.	PS51963	BCOV_NSP1_C	148	179	Beta coronavirus (BetaCoV) Nsp1 C-terminal domain profile
3.	PS51943	COV_NSP3A_UBL	820	929	Coronavirus Nsp3a Ubl domain profile
4.	PS51154	MACRO	1025	1194	Macro domain profile
5.	PS51940	SARS_NSP3C_N	1231	1359	Sarbecovirus Nsp3c-N domain profile
6.	PS51941	BCOV_NSP3C_M	1367	1494	Betacoronavirus Nsp3c-M domain profile
7.	PS51942	BCOV_NSP3C_C	1496	1561	Betacoronavirus Nsp3c-C domain profile
8.	PS51944	COV_NSP3D_UBL	1565	1620	Coronavirus Nsp3d Ubl domain profile
9.	PS51124	PEPTIDASE_C16	1634	1898	Peptidase family C16 domain profile
10.	PS51945	BCOV_NSP3E_NAB	1911	2021	Betacoronavirus Nsp3e nucleic acid-binding (NAB) domain profile
11.	PS51946	COV_NSP4C	3165	3263	Coronavirus Nsp4 C-terminal (Nsp4C) domain profile
12.	PS51442	M_PRO	3264	3569	Coronavirus main protease (M-pro) domain profile
13.	PS51949	COV_NSP7	3860	3942	Coronavirus RNA-dependent RNA polymerase (RdRp) Nsp7 cofactor domain profile
14.	PS51950	COV_NSP8	3943	4140	Coronavirus RNA-dependent RNA polymerase (RdRp) Nsp8 cofactor domain profile
15.	PS51951	COV_NSP9_SSRNA_BD	4141	4253	Coronavirus Nsp9 single-stranded RNA (ssRNA)-binding domain profile
16.	PS51952	COV_EXON_MTASE_COACT	4254	4392	Coronavirus (CoV) ExoN/MTase coactivator domain profile

**Fig. 3** The post-translational modifications prediction like phosphorylation, methylation, glycosylation and sumoylation in the protein Mpro

significantly (Mishra et al. 2021). Recently, ground-breaking findings of SARS-CoV-2 and host protein inter-proteomic PTMs as well as their critical roles in virus infection and pathogenesis have been observed (Bouhaddou et al. 2020). SARS-CoV-2 is believed to be able to infect host cells, escape host immune defence, multiply within host cells and move from one host to another because of the variability induced by inter-proteomic PTMs of the virus and host proteins.

Phosphoproteomic investigation on serine, threonine, and tyrosine alterations in bacteria and viruses revealed that phosphorylated threonine was more prevalent than serine and tyrosine (Macek et al. 2007). In the Mpro protein, more phosphorylated threonine has been predicted than serine and tyrosine residues, as expected. In addition, the existence of three putative methylation and three sumoylation sites raised the possibility of functional changes in the protein. The tetrapeptide motif *B-K-x-D/E* is found in most SUMO-modified proteins, where *B* is a hydrophobic residue, *K* is the SUMO-conjugated lysine, *x* is any amino acid (aa), and *D* or *E* is an acidic residue. Even though predicting the role of each PTM in the protein is difficult and necessitates extensive molecular techniques. The possibilities of changes in Mpro protein exhibit additional orders of peptidase function regulation. These preliminary findings will be beneficial in the future to inspire a more profound knowledge of the Mpro protein mechanism of action and to offer a foundation for additional antiviral drugs targeting research (Fung and Liu 2018).

Mpro, a non-structural protein encoded by orf1a, was also suggested as a possible antigen (Can et al. 2020). Using a threshold value of 0.4 for antigenicity prediction, it was determined that Mpro total protective antigen prediction score was 0.4159 (Probable Antigen). Hence, Mpro encoded by orf1a was predicted to be a potential antigen



for SARS-CoV-2. Despite the fact that there was no significant difference in anticipated antigenicity values for likely vaccination candidate proteins. Furthermore, the effect of a point mutation on the Mpro protein was seen during heatmap generation.

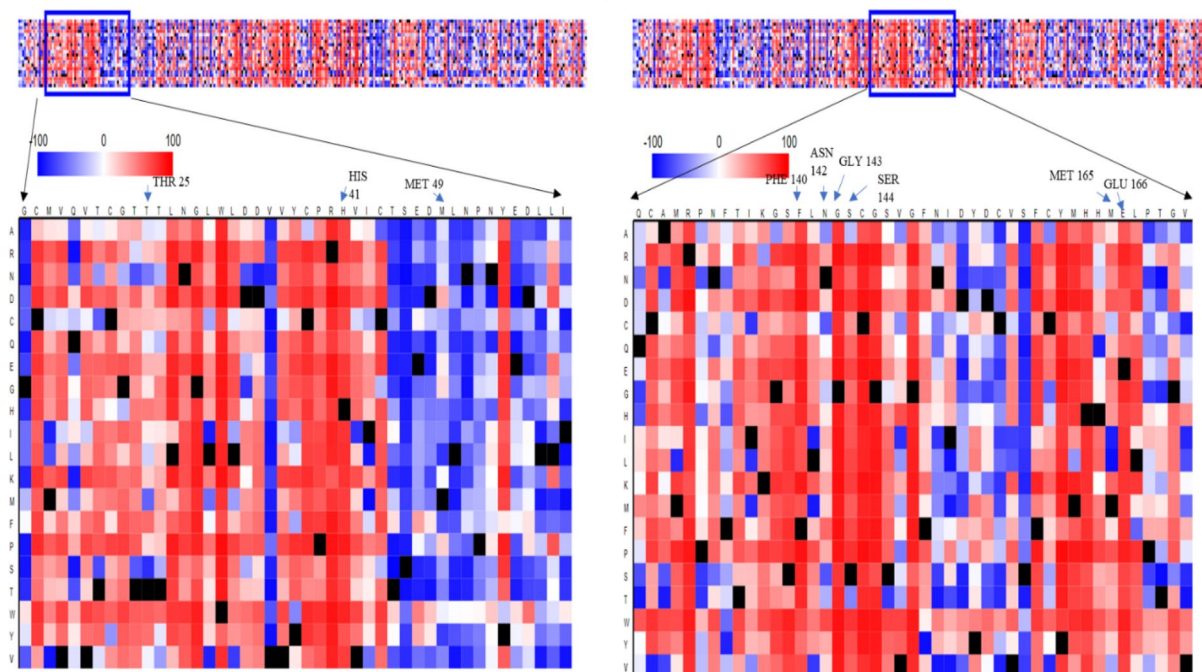
In a heatmap representation, we show each substitution separately for each protein location. “A high score (score > 50, strong signal for effect) is shown in dark red, a weak signal (− 50 score 50), and a low score (score − 50, strong signal for neutral/no effect) is shown in blue” (Yachdav et al. 2014). The wild-type residues are highlighted in black. In the above heatmap representation, we have targeted mainly active sites of Mpro protein residues. It was observed that amino acid residues HIS<sup>41</sup>, PHE<sup>140</sup>, GLY<sup>143</sup>, SER<sup>144</sup>, and GLU<sup>166</sup> showed a strong signal for effect. Whereas THR<sup>25</sup>, MET<sup>49</sup>, ASN<sup>142</sup>, and MET<sup>165</sup> show strong signals for neutral/no effect (Fig. 4).

### Tertiary structure analysis, validation, the secondary structure arrangement, and topology

For the tertiary structure of Mpro, a protein selection similarity search was performed using PDB BLAST, and the best resolution containing structure was selected. PyMOL (Schrödinger Inc.) was used to view the model's publishable picture (Fig. 5a) (DeLano 2002). Structure assessment was the most important part of structure prediction. for the

assessment of tertiary protein structure, ProSA and VADAR server tools were used. The ProSA server calculates the energy while considering the atomic pair potential and the solvent exposure to protein residues (Sippl 1995). The Mpro protein 'Z' score is − 7.34, which is within the range. Our 'Z' score is identical to the 'Z' score acquired from experimental analysis (NMR and X-ray crystallography). The majority of the Mpro amino acid is found in the negative quadrant of the generated energy plot (Fig. 5c, d). The negative energy of all amino acid residues shows that our model is of high quality in the energy plot. In addition, the VADAR analysis prediction revealed 72 (23%) helix, 95 (31%) beta, 139 (45%) coil, and 88 (28%) turns with mean H bond energy of − 1.8 (SD=1.0) against the expected value of − 2.0 (SD=0.8) in the protein model. As measured by several servers, these parameters indicated that the model was efficient and structural stability was maintained, thereby validating the structure. The PDBsum tool was used to analyze the secondary structural elements more intensively. Here, the amino acid sequence of Mpro is the secondary structure displaying 10 helices (H1–H10), while  $\beta$ -sheet motifs are composed of 13  $\beta$ -strands, 26 betas, and 3 gamma turns.

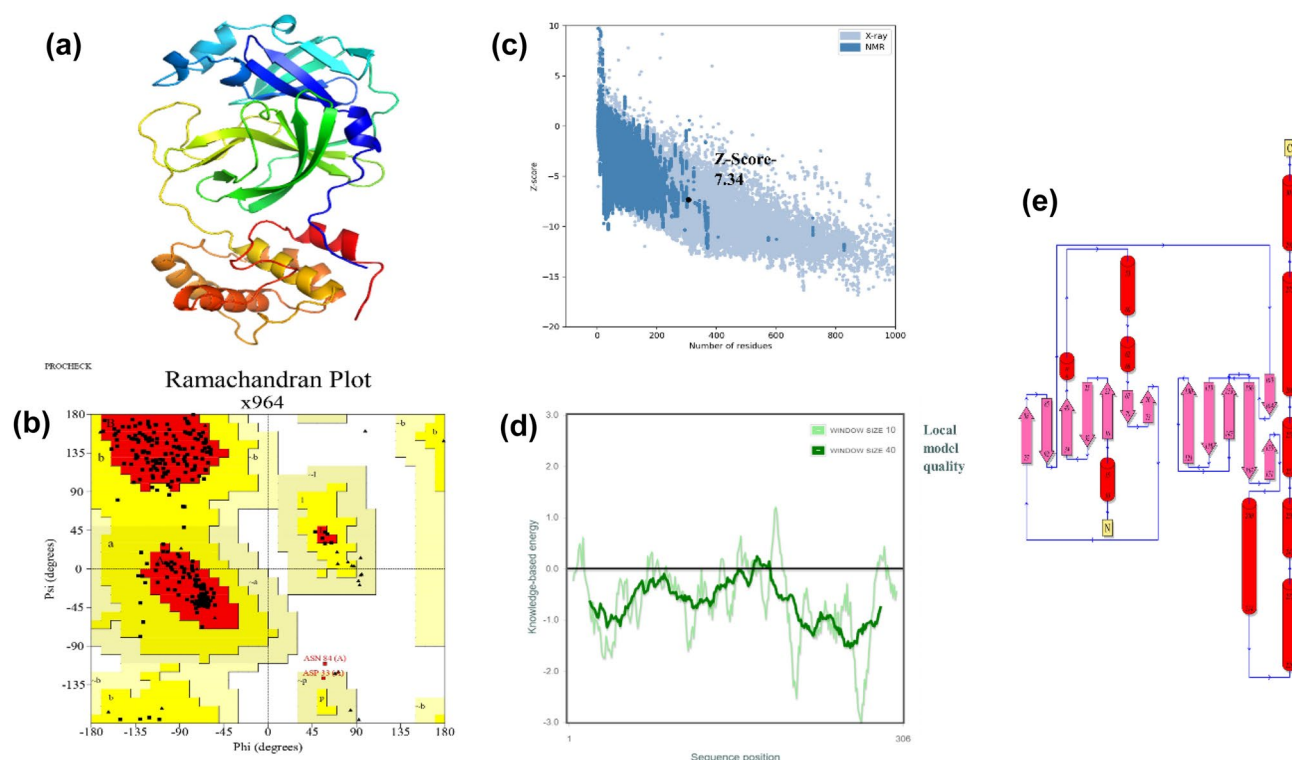
Thereafter, Mpro was structurally validated by PROCHECK; according to an Analysis of the Ramachandran plot, 90.6% of residues were found in the most favoured locations, 8.7 percent in additional permitted regions, 0.4% in generously allowed regions, and 0.4 percent in disallowed



**Fig. 4** A high score (score > 50, strong signal for effect) is denoted by dark red, a weak signal (− 50 score 50), and a low score (score − 50, strong signal for neutral/no impact) is marked by blue. The wild-type

residues are highlighted in black. The zoom panel below the main panel is a blowout of a specific region in the heatmap presented above and contained in the blue frame





**Fig. 5** Mpro model's quality assessment **a** The 3D structure of the Mpro protein, which was based on the PDB template (PDB ID: 6LU7). The model was visualized using PyMOL **b** Ramachandran plot was used for the validation of the Mpro model. **c** ProSA-web Z-scores in terms of length obtained by NMR spectroscopy (light

blue) with X-ray crystallography (dark blue) for all polypeptide chains in the PDB. A massive black dot represents Mpro's Z-score in that range. **d** Energy map of all protein residues from the ProSA server. **e** Helices were depicted as cylinders, while strands were depicted as arrows

regions (Fig. 5b). It is possible that the presence of fewer proline residues in the peptide chain could be the cause behind it (Lovell et al. 2003). Its topology was depicted in (Fig. 5e), which demonstrated how the helices and strands of the protein were arranged and connected. The protein chain included two domains, a C-terminal domain and an N-terminal domain were both folded into a mixed/structure.

## Conclusion

The above in-silico characterization of Mpro protein was identified to be a 306 amino acid acidic, non-polar, and heat-stable globular protein. (PDB ID:6LU7). The Mpro protein included 4.24% proline residues, which resulted in a less conformational rigidity in the protein structure due to the lower proline content. The results of the sequence and phylogenetic analyses were as follows: in a phylogenetic tree, the tested Mpro proteins homologs form two clusters. Sequence and phylogenetic analysis revealed the evolutionary conservation of these functional motifs, especially in the Peptidase-C30 domain region. M-pro is a dimer with three different domains, I, II, and III, in each subunit. The helical

domains III as well as the surface loops are the most diverse regions. The cleft between domains I and II are highly conserved across all coronaviruses, implying that antiviral drugs targeting this pocket should be effective against all coronaviruses. It was observed that 15 motifs out of 30 are fully conserved, and the majority of them revealed their identity as the Peptidase-C30 domains. The motifs related to active sites of proteins are fully conserved. Further, it was observed that closely related organisms have the majority of motifs common in them. Many post-translational modifications (PTMs) were found, which may result in changes in Mpro protein that exhibit additional orders of peptidase function regulation. The biological impact of PTMs on several coronavirus proteins has yet to be fully defined, and further research integrating reverse genetics and appropriate in vivo models is required. Mpro protein was known to elicit a significant immunological response, according to the findings of these studies. It was observed that some amino acid residues of active sites show strong signals for the effect of point mutation. The Mpro protein's physicochemical properties, as well as its secondary arrangement, tertiary structure, validation, and homologs evolution, were predicted using in-silico research.

**Supplementary Information** The online version contains supplementary material available at <https://doi.org/10.1007/s42485-023-00105-9>.

**Acknowledgements** A. Gupta (09/013(0912)/2019-EMR-I) and N. Sahu (09/013(0927)/2020-EMR-I) are thankful to the Council of Scientific and Industrial Research (CSIR), New Delhi, India, for the Senior Research Fellowship (SRF) and Junior Research Fellowship (JRF), respectively. All authors would also like to take this opportunity to thank the Head of the Department of Botany and the Coordinator, School of Biotechnology, Banaras Hindu University, for providing the necessary facilities and encouragement to carry out this work.

**Author contributions** AG and RPS conceptualized the aims and methodology. AG performed formal data analysis, was involved in data curation, and produced the first draft of the manuscript under the supervision of RPS. NS Performed formal data analysis and was involved in data curation. VKS performed formal data analysis and validation. All authors contributed to editing and reviewing the manuscript.

**Funding** This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

**Data availability** The data of our research can be acquired from the supplementary materials uploaded with this article.

## Declarations

**Conflict of interest** The authors have no conflict of interest to declare that are relevant to the content of this article.

## References

- Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. *J Mol Biol* 215(3):403–410
- Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucl Acids Res* 25(17):3389–3402
- Arnold K, Bordoli L, Kopp J, Schwede T (2006) The SWISS-MODEL workspace: a web-based environment for protein structure homology modelling. *Bioinformatics* 22(2):195–201
- Bailey TL, Johnson J, Grant CE, Noble WS (2015) The MEME suite. *Nucl Acids Res* 43(W1):W39–W49
- Bharadwaj S, Lee KE, Dwivedi VD, Kang SG (2020) Computational insights into tetracyclines as inhibitors against SARS-CoV-2 Mpro via combinatorial molecular simulation calculations. *Life Sci* 257:118080
- Bouhaddou M, Memon D, Meyer B, White KM, Rezeli VV, Marrero MC, Polacco BJ, Melnyk JE, Ulferts S, Kaake RM, Batra J (2020) The global phosphorylation landscape of SARS-CoV-2 infection. *Cell* 182(3):685–712
- Cain JA, Solis N, Cordwell SJ (2014) Beyond gene expression: the impact of protein post-translational modifications in bacteria. *J Proteom* 97:265–286
- Can H, Köseoglu AE, Erkunt Alak S, Güvendi M, Döşkaya M, Karakavuk M, Gürüz AY, Ün C (2020) In silico discovery of antigenic proteins and epitopes of SARS-CoV-2 for the development of a vaccine or a diagnostic approach for COVID-19. *Sci Rep* 10(1):1–16
- Chakraborty C, Sharma AR, Bhattacharya M, Sharma G, Lee SS (2020) The 2019 novel coronavirus disease (COVID-19) pandemic: a zoonotic prospective. *Asian Pac J Trop Med* 13(6):242
- Dai W, Zhang B, Jiang XM, Su H, Li J, Zhao Y, Xie X, Jin Z, Peng J, Liu F, Li C (2020) Structure-based design of antiviral drug candidates targeting the SARS-CoV-2 main protease. *Science* 368(6497):1331–1335
- DeLano WL (2002) The PyMOL molecular graphics system. DeLano Scientific, San Carlos
- Drosten C, Günther S, Preiser W, Van Der Werf S, Brodt HR, Becker S, Rabenau H, Panning M, Kolesnikova L, Fouchier RA, Berger A (2003) Identification of a novel coronavirus in patients with severe acute respiratory syndrome. *NEJM* 348(20):1967–1976
- Fung TS, Liu DX (2018) Post-translational modifications of coronavirus proteins: roles and function. *Fut Virol* 13(6):405–430
- Gasteiger E, Hoogland C, Gattiker A, Wilkins MR, Appel RD, Bairoch A (2005) Protein identification and analysis tools on the ExPASy server. In: *The Proteomics Protocols Handbook*, pp 571–607
- Ghosh R, Chakraborty A, Biswas A, Chowdhuri S (2021) Identification of polyphenols from *Broussonetia papyrifera* as SARS CoV-2 main protease inhibitors using in silico docking and molecular dynamics simulation approaches. *J Biomol Struct Dyn* 39(17):6747–6760
- Gupta A, Sahu N, Singh AP, Singh VK, Singh SC, Upadhye VJ, Mathew AT, Kumar R, Sinha RP (2022) Exploration of novel lichen compounds as inhibitors of SARS-CoV-2 Mpro: ligand-based design, molecular dynamics, and ADMET analyses. *Appl Biochem Biotechnol* 194(12):6386–6406
- Hegyí A, Ziebuhr J (2002) Conservation of substrate specificities among coronavirus main proteases. *J Gen Virol* 83(3):595–599
- Jiménez-Alberto A, Ribas-Aparicio RM, Aparicio-Ozores G, Castelán-Vega JA (2020) Virtual screening of approved drugs as potential SARS-CoV-2 main protease inhibitors. *Comput Biol Chem* 88:107325
- Jin Z, Du X, Xu Y, Deng Y, Liu M, Zhao Y, Zhang B, Li X, Zhang L, Peng C, Duan Y (2020) Structure of Mpro from SARS-CoV-2 and discovery of its inhibitors. *Nature* 582(7811):289–293
- Joshi T, Sharma P, Joshi T, Pundir H, Mathpal S, Chandra S (2021) Structure-based screening of novel lichen compounds against SARS Coronavirus main protease (Mpro) as potential inhibitors of COVID-19. *Mol Divers* 25:1665–1677
- Kharwar S, Bhattacharjee S, Mishra AK (2021) Disentangling the impact of sulfur limitation on exopolysaccharide and functionality of Alr2882 by in silico approaches in *Anabaena* sp. *PCC 7120*. *Appl Biochem Biotechnol* 193:1447–1468
- Kim Y, Liu H, Galasiti Kankanamalage AC, Weerasekara S, Hua DH, Groutas WC, Chang KO, Pedersen NC (2016) Reversal of the progression of fatal coronavirus infection in cats by a broad-spectrum coronavirus protease inhibitor. *PLoS Pathog* 12(3):1005531
- Kneller DW, Phillips G, O'Neill HM, Jedrzejczak R, Stols L, Langan P, Joachimiak A, Coates L, Kovalevsky A (2020) Structural plasticity of SARS-CoV-2 3CL Mpro active site cavity revealed by room temperature X-ray crystallography. *Nat Commun* 11(1):16
- Kumar S, Stecher G, Li M, Knyaz C, Tamura K (2018) MEGA X: molecular evolutionary genetics analysis across computing platforms. *Mol Biol Evol* 35(6):1547
- Lovell SC, Davis IW, Arendall WB III, De Bakker PI, Word JM, Prisant MG, Richardson JS, Richardson DC (2003) Structure validation by  $\alpha$  geometry:  $\phi$ ,  $\psi$  and  $C\beta$  deviation. *Proteins Struct Funct Genet* 50(3):437–450
- Macek B, Mijakovic I, Olsen JV, Gnad F, Kumar C, Jensen PR, Mann M (2007) The serine/threonine/tyrosine phosphoproteome of the model bacterium *Bacillus subtilis*. *Mol Cell Proteom* 6(4):697–707
- Mirza MU, Froeyen M (2020) Structural elucidation of SARS-CoV-2 vital proteins: Computational methods reveal potential drug candidates against main protease, Nsp12 polymerase and Nsp13 helicase. *J Pharm Anal* 10(4):320–328

- Mishra S, Bassi G, Nyomba BG (2021) Inter-proteomic posttranslational modifications of the SARS-CoV-2 and the host proteins—a new frontier. *Exp Biol Med* 246(7):749–757
- Rothan HA, Byrareddy SN (2020) The epidemiology and pathogenesis of coronavirus disease (COVID-19) outbreak. *J Autoimmune* 109:102433
- Sigrist CJ, De Castro E, Cerutti L, Cuche BA, Hulo N, Bridge A, Bougueleret L, Xenarios I (2012) New and continuing developments at PROSITE. *Nucl Acids Res* 41(D1):D344–D347
- Singh S, Biswas S, Srivastava A, Mishra Y, Chaturvedi TP (2021) In silico characterization and structural modeling of a homeobox protein MSX1 from *Homo sapiens*. *Inform Med Unlocked* 22:100497
- Sinha AU, Meller J (2007) Cinteny: flexible analysis and visualization of synteny and genome rearrangements in multiple organisms. *BMC Bioinform* 8(1):1–9
- Sinha AU, Meller J (2008) Sensitivity analysis for reversal distance and breakpoint reuse in genome rearrangements. In: *Biocomputing*, pp 37–48
- Sippl MJ (1993) Recognition of errors in three-dimensional structures of proteins. *Proteins Struct Funct Genet* 17:355–362
- Sippl MJ (1995) Knowledge-based potentials for proteins. *Curr Opin Struct Biol* 5:229–235
- Szymanski CM, Burr DH, Guerry P (2002) Campylobacter protein glycosylation affects host cell interactions. *Infect Immun* 70(4):2242–2244
- Tian W, Chen C, Liang J (2018) CASTp 3.0: computed atlas of surface topography of proteins and beyond. *Biophys J* 114(3):50a
- Walls AC, Park YJ, Tortorici MA, Wall A, McGuire AT, Veasler D (2020) Structure, function, and antigenicity of the SARS-CoV-2 spike glycoprotein. *Cell* 181(2):281–292
- Waterhouse AM, Procter JB, Martin DM, Clamp M, Barton GJ (2009) Jalview version 2—a multiple sequence alignment editor and analysis workbench. *Bioinformatics* 25(9):1189–1191
- Willard L, Ranjan A, Zhang H, Monzavi H, Boyko RF, Sykes BD, Wishart DS (2003) VADAR: a web server for quantitative evaluation of protein structure quality. *Nucl Acids Res* 31(13):3316–3319
- Xue Y, Zhou F, Fu C, Xu Y, Yao X (2006) SUMOsp: a web server for sumoylation site prediction. *Nucl Acids Res* 34:W254–W257
- Yachdav G, Kloppmann E, Kajan L, Hecht M, Goldberg T, Hamp T, Hönigschmid P, Schafferhans A, Roos M, Bernhofer M, Richter L (2014) PredictProtein—an open resource for online prediction of protein structural and functional features. *Nucl Acids Res* 42(W1):W337–W343
- Yoshimoto FK (2020) The proteins of severe acute respiratory syndrome coronavirus-2 (SARS CoV-2 or n-COV19), the cause of COVID-19. *The Protein J* 39:198–216
- Yu H, Zhao Y, Guo C, Gan Y, Huang H (2015) The role of proline substitutions within flexible regions on thermostability of luciferase Proteins and proteomics. *BBA Proteins Proteomics* 1854:65–72
- Zaki AM, Van Boheemen S, Bestebroer TM, Osterhaus AD, Fouchier RA (2012) Isolation of a novel coronavirus from a man with pneumonia in Saudi Arabia. *NEJM* 367(19):1814–1820
- Zhang J, Wang Y (2021) Bilobetin, a novel small molecule inhibitor targeting influenza virus polymerase acidic (PA) endonuclease was screened from plant extracts. *Nat Prod Res* 2021:1–4
- Zhang CC, Jang J, Sakr S, Wang L (2005) Protein phosphorylation on Ser, Thr and Tyr residues in cyanobacteria. *Microb Physiol* 9(3–4):154–166
- Zhang L, Lin D, Sun X, Curth U, Drosten C, Sauerhering L, Becker S, Rox K, Hilgenfeld R (2020) Crystal structure of SARS-CoV-2 main protease provides a basis for design of improved  $\alpha$ -ketoamide inhibitors. *Science* 368:64–89
- Zhao Q, Li S, Xue F, Zou Y, Chen C, Bartlam M, Rao Z (2008) Structure of the main protease from a global infectious human coronavirus, HCoV-HKU1. *J Virol* 82(17):8647–8655
- Zhao W, Zhong Z, Xie X, Yu Q, Liu J (2020) Relation between chest CT findings and clinical conditions of coronavirus disease (COVID-19) pneumonia: a multicenter study. *AJR Am J Roentgenol* 214(5):1072–1077
- Zhou P, Yang XL, Wang XG, Hu B, Zhang L, Zhang W, Si HR, Zhu Y, Li B, Huang CL, Chen HD (2020) A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature* 579(7798):270–273

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.