Setayesh Yazdani

M.Sc. Student at Structural Genomics Consortium (SGC) Toronto

Pharmacology and Toxicology Department, University of Toronto

****

"Methods of Mapping the genetic variations of SARS-CoV-2 onto its proteins' crystal structures"

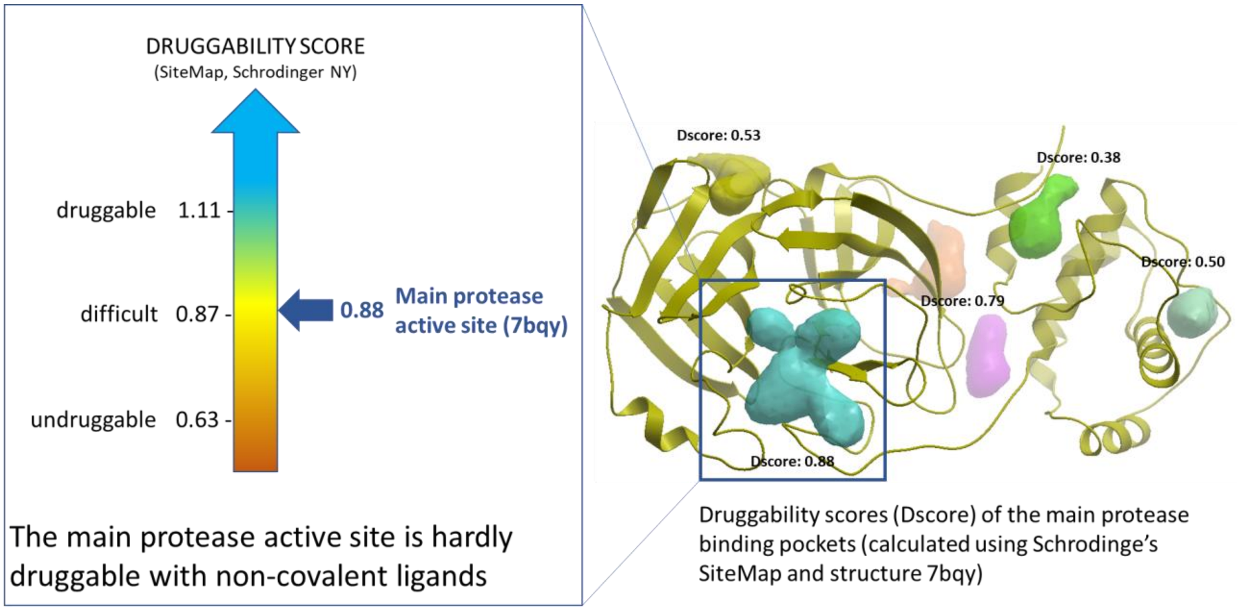
****

The relevant ICM scripts can be found in step-wise_scritps.icm file in this Zenodo post.

Acronym for the main protease: MPro

Step 1: Identifying druggable binding sites on SARS-CoV-2 MPro crystal structure.

We used the PocketFinder function in ICM (Molsoft, San Diego) to find potential druggable binding sites using the PDB structure 7bqy as shown in Figure 1. Druggability analysis of the pockets on 7bqy was done using SiteMap(Schrodinger, NY). The druggability score (D-score) of the catalytic pocket was 0.88.



The main protease active site is hardly druggable with non-covalent ligands

Druggability scores (Dscore) of the main protease binding pockets (calculated using Schrodinge's SiteMap and structure 7bqy)

Figure 1. Druggability analysis of binding pockets at the surface of the main protease. Druggability scores (Dscore) are calculated with SiteMap (Schrodinger, NY)

Step 2: Defining the neighboring residues of MPro catalytic site.

We used the structure of SARS-CoV-2 MPro bound to ligand N3 (PDB code 7bqy). After manually checking the residues with sidechain atoms within 2.8 Å vicinity of the wall of the catalytic pocket as defined by ICM, we excluded the amino acid residues with sidechains pointing away from the pocket rather than towards the pocket. We identified 21 amino acid residues, which we refer to as neigh_res. a_7bqy_1.a refers to the structure of MPro. The command below summarizes our selection of amino acid residues of the catalytic site which is also highlighted in blue in Figure 2.

neigh_res=a_7bqy_1.a/25,27,41,44,49,52,54,140,142,144:145,163:168,172,187,189,192

The full list of neigh_res include: Thr25, Leu27, His41, Cys44, Met49, Pro52, Tyr54, Phe140, Asn142, Ser144, Cys145, His163, His164, Met165, Glu166, Leu167, Pro168, His172, Asp187, Gln189, Gln192.

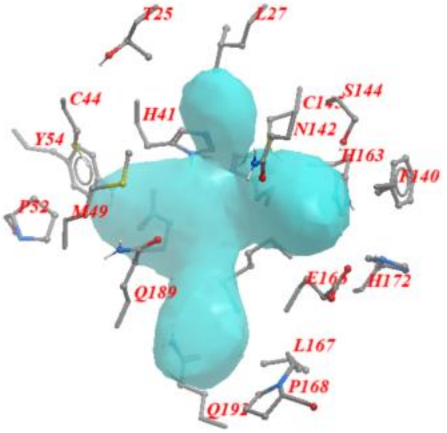
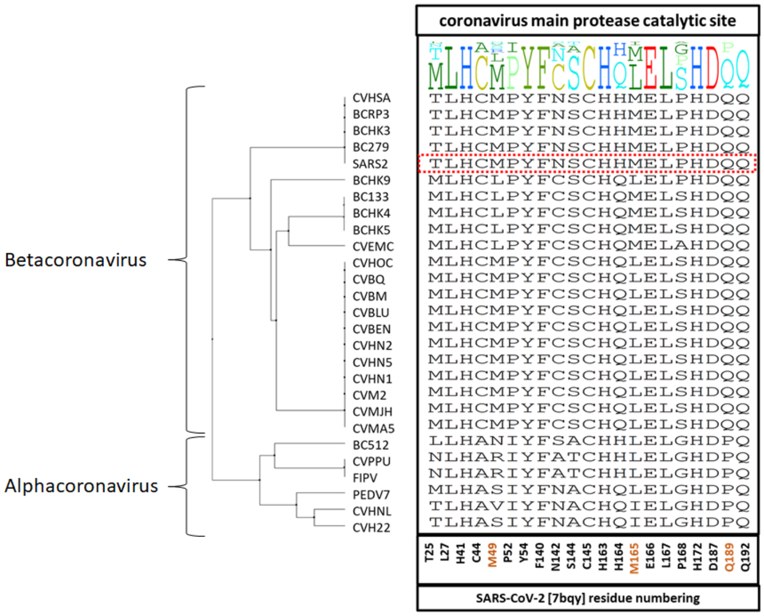


Figure 2. The neighbouring amino acid residues of the MPro catalytic site (PDB code: 7bqy).

Step 3: Downloading sequences from UniProt website for SARS-COV-2 MPro sequence.

Over 400 reviewed protein sequences of alpha and betacoronaviruses were downloaded from UniProt. (https://www.uniprot.org/). Using the SARS-CoV-2 MPro sequence, we searched for sequences from UniProt database. We found 27 unique sequences with different percentage sequence identity values. We then made a diversity dendrogram while looking at the MPro catalytic site as the selection block (21 residues lining the catalytic site) across these 27 sequences. The diversity dendrogram is shown in Figure 3. The sequences are divided into two groups: alpha and beta coronavirus. The consensus profile is shown using colored symbolic letters with varying sizes, which reflect residue conservation. We have used a red dashed-line rectangular shape to highlight the SARS-CoV-2 MPro catalytic site. Lastly, in orange, are residues that were non-conserved but were deemed critical in terms of interaction with ligand/inhibitor in the binding site according to PDB structure 7bqy.


coronavirus main protease catalytic site
Betacoronavirus
Alphacoronavirus
CVHSA TLHCMPYFNSCHHMELPHDQQ
BCRP3 TLHCMPYFNSCHHMELPHDQQ
BCHK3 TLHCMPYFNSCHHMELPHDQQ
BC279 TLHCMPYFNSCHHMELPHDQQ
SARS2 TLHCMPYFNSCHHMELPHDQQ
BCHK9 MLHCLPYFCSCHQLELPHDQQ
BC133 MLHCLPYFCSCHQMELSHDQQ
BCHK4 MLHCLPYFCSCHQMELSHDQQ
BCHK5 MLHCLPYFCSCHQMELSHDQQ
CVEMC MLHCLPYFCSCHQMELAHDQQ
CVHOC MLHCMPYFCSCHQLELSHDQQ
CVBQ MLHCMPYFCSCHQLELSHDQQ
CVBM MLHCMPYFCSCHQLELSHDQQ
CVBLU MLHCMPYFCSCHQLELSHDQQ
CVBEN MLHCMPYFCSCHQLELSHDQQ
CVHN2 MLHCMPYFCSCHQLELSHDQQ
CVHN5 MLHCMPYFCSCHQLELSHDQQ
CVHN1 MLHCMPYFCSCHQLELSHDQQ
CVM2 MLHCMPYFCSCHQLELSHDQQ
CVMJH MLHCMPYFCSCHQLELSHDQQ
CVMA5 MLHCMPYFCSCHQLELSHDQQ
BC512 LLHANIYFSACHHLELGHDPQ
CVPPU NLHARIYFATCHHLELGHDPQ
FIPV NLHARIYFATCHHLELGHDPQ
PEDV7 MLHASIYFNACHQLELGHDPQ
CVHNL TLHAVIYFNACHQIELGHDPQ
CVH22 TLHASIYFNACHQIELGHDPQ
T25 L27 H41 C44 M49 P52 Y54 F140 N142 S144 C145 H163 H164 M165 E166 L167 P168 H172 D187 Q189 Q192
SARS-CoV-2 [7bqy] residue numbering

Entry	Organism
CVHSA	Human SARS coronavirus (SARS-CoV) (Severe acute respiratory syndrome coronavirus)
BCRP3	Bat coronavirus Rp3/2004 (BtCoV/Rp3/2004) (SARS-like coronavirus Rp3)
BCHK3	Bat coronavirus HKU3 (BtCoV) (SARS-like coronavirus HKU3)
BC279	Bat coronavirus 279/2005 (BtCoV) (BtCoV/279/2005)
SARS2	Severe acute respiratory syndrome coronavirus 2 (2019-nCoV) (SARS-CoV-2)
BCHK9	Bat coronavirus HKU9 (BtCoV) (BtCoV/HKU9)
BC133	Bat coronavirus 133/2005 (BtCoV) (BtCoV/133/2005)
BCHK4	Bat coronavirus HKU4 (BtCoV) (BtCoV/HKU4/2004)
BCHK5	Bat coronavirus HKU5 (BtCoV) (BtCoV/HKU5/2004)
CVEMC	Middle East respiratory syndrome-related coronavirus (Human coronavirus EMC)
CVHOC	Human coronavirus OC43 (HCoV-OC43)
CVBQ	Bovine coronavirus (strain Quebec) (BCoV) (BCV)
CVBM	Bovine coronavirus (strain Mebus) (BCoV) (BCV)
CVBLU	Bovine coronavirus (strain 98TXSF-110-LUN) (BCoV-LUN) (BCV)
CVBEN	Bovine coronavirus (strain 98TXSF-110-ENT) (BCoV-ENT) (BCV)
CVHN2	Human coronavirus HKU1 (isolate N2) (HCoV-HKU1)
CVHN5	Human coronavirus HKU1 (isolate N5) (HCoV-HKU1)
CVHN1	Human coronavirus HKU1 (isolate N1) (HCoV-HKU1)
CVM2	Murine coronavirus (strain 2) (MHV-2) (Murine hepatitis virus)
CVMJH	Murine coronavirus (strain JHM) (MHV-JHM) (Murine hepatitis virus)
CVMA5	Murine coronavirus (strain A59) (MHV-A59) (Murine hepatitis virus)
BC512	Bat coronavirus 512/2005 (BtCoV) (BtCoV/512/2005)
CVPPU	Porcine transmissible gastroenteritis coronavirus (strain Purdue) (TGEV)
FIPV	Feline coronavirus (strain FIPV WSU-79/1146) (FCoV)
PEDV7	Porcine epidemic diarrhea virus (strain CV777) (PEDV)
CVHNL	Human coronavirus NL63 (HCoV-NL63)
CVH22	Human coronavirus 229E (HCoV-229E)

Figure 3. The diversity dendrogram of the coronavirus main protease catalytic site. 27 entries from the UniProt database and their corresponding organism names are shown in the table. Critical non-conserved residues are highlighted in orange.

Step 4: Manual assessments of SARS-Cov-2, SARS-CoV and MERS MPro's structures in complex with their inhibitors.

We selected three MPro structures from the PDB: SARS-Cov-2(PDB code 7bqy), SARS-CoV (PDB code 2gx4) and MERS (pdb code 5wkj). We first visually and manually assessed these structures in terms of their interactions with their inhibitors in the catalytic site. Second, we highlighted if we thought any conserved residues were critical for those interactions. For all these MPro examples, we used ICM function icmPocketFinder to locate the catalytic pocket. For all of these structural analyses, we looked at the 21 residues lining the catalytic site. We created 2-dimensional (2D) ligand-protein interactions using the MPro's crystal structures and their corresponding ligand (inhibitors) in Proteins.plus PoseView (https://proteins.plus/). The following Figures 4, 5, and 6 show the catalytic side pockets and ligands positioning based on the available crystal structures. We highlighted the critical non-conserved residue in orange.

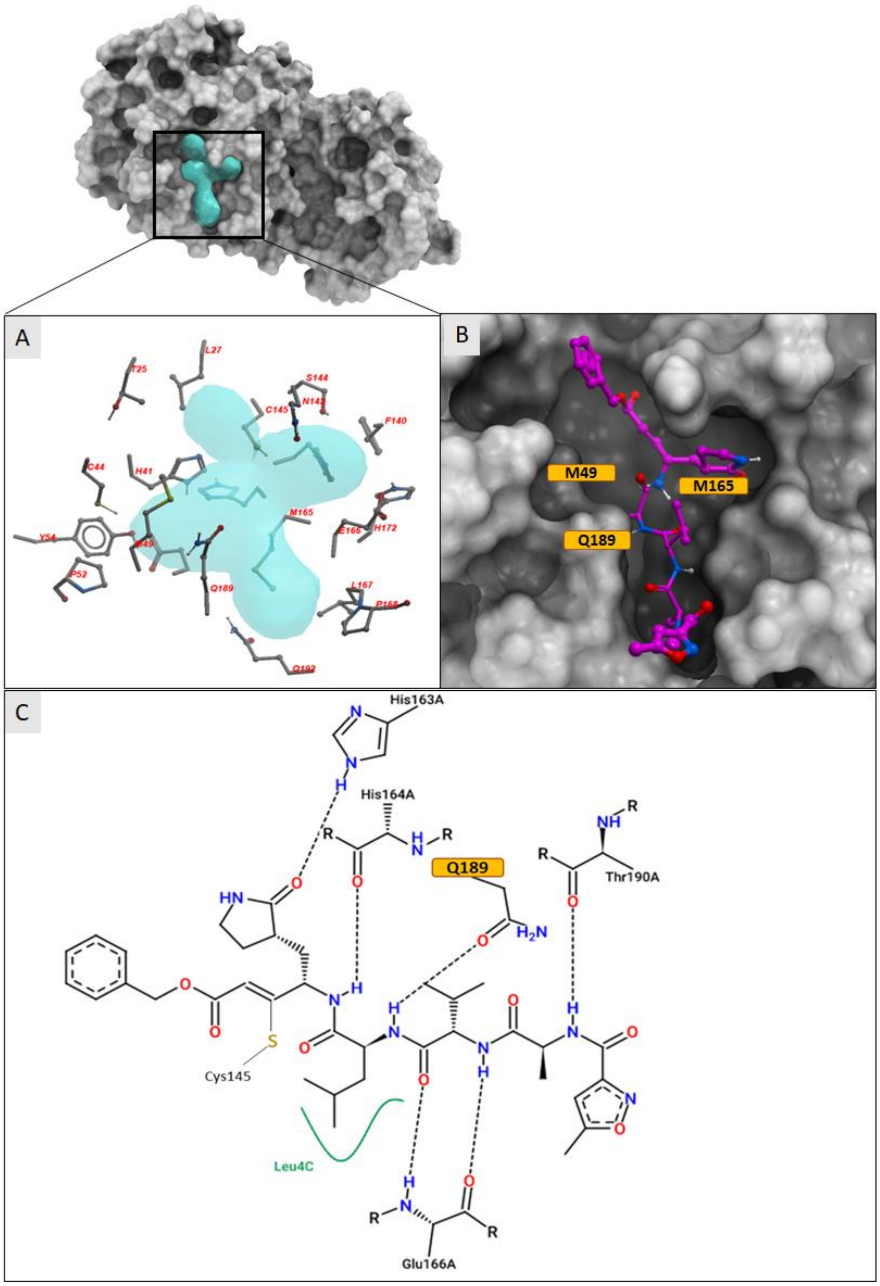


Figure 4. (A) The catalytic site of SARS-CoV-2 main protease (PDB code 7bqy). **(B)** Non-conserved critical sidechains are highlighted in orange. Crystallized inhibitor: purple. **(C)** 2D interaction diagram of a SARS-CoV-2 main protease inhibitor (created using Proteins.plus PoseView, PDB code 7bqy). A non-conserved critical sidechain is highlighted in orange.

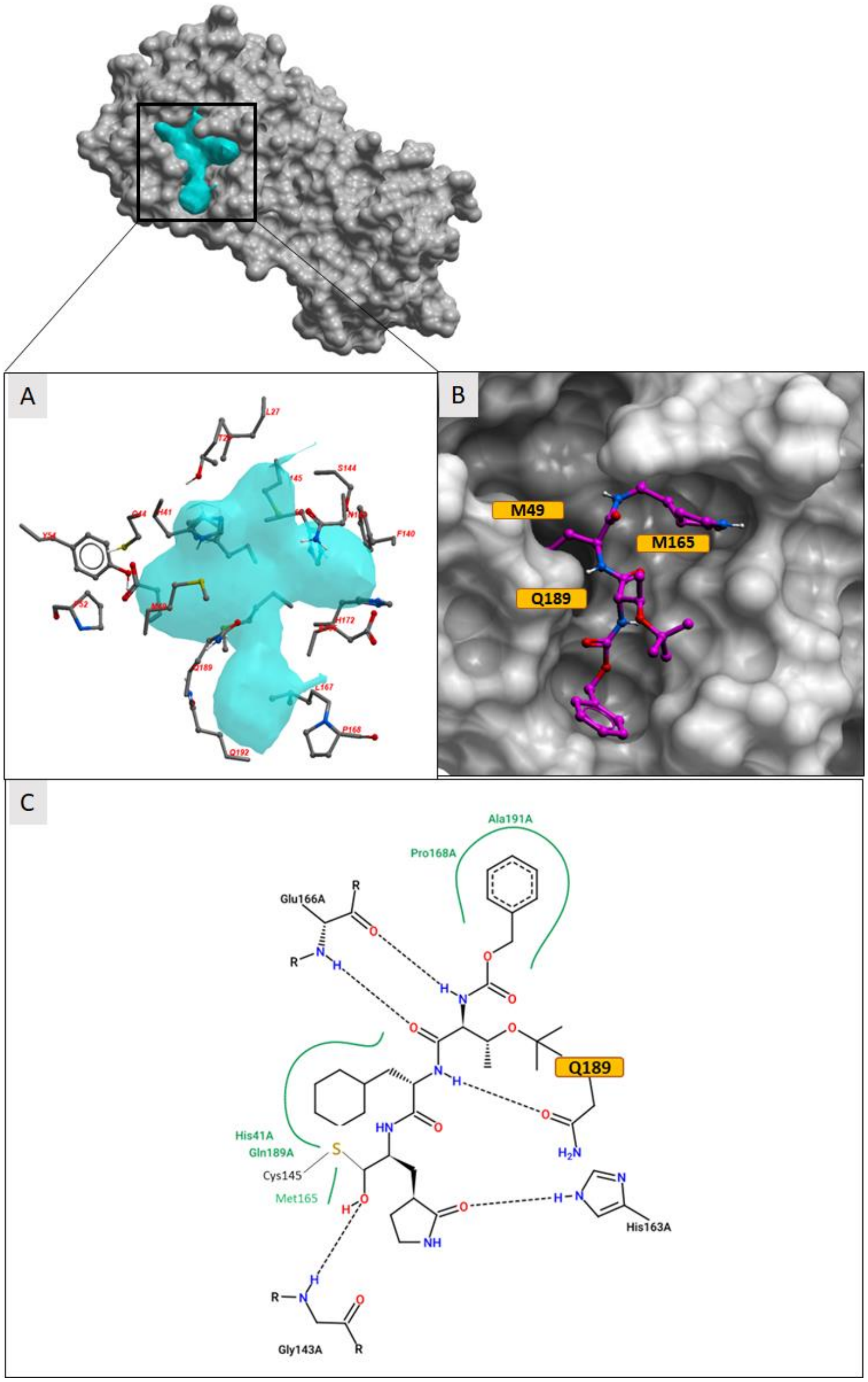


Figure 5. (A) The catalytic site of SARS-CoV main protease (PDB code 2gx4). **(B)** Non-conserved critical sidechains are highlighted in orange. Crystallized inhibitor: purple. **(C)** 2D interaction diagram of a SARS-CoV main protease inhibitor (created using Proteins.plus PoseView, PDB code 2gx4). A non-conserved critical sidechain is highlighted in orange.

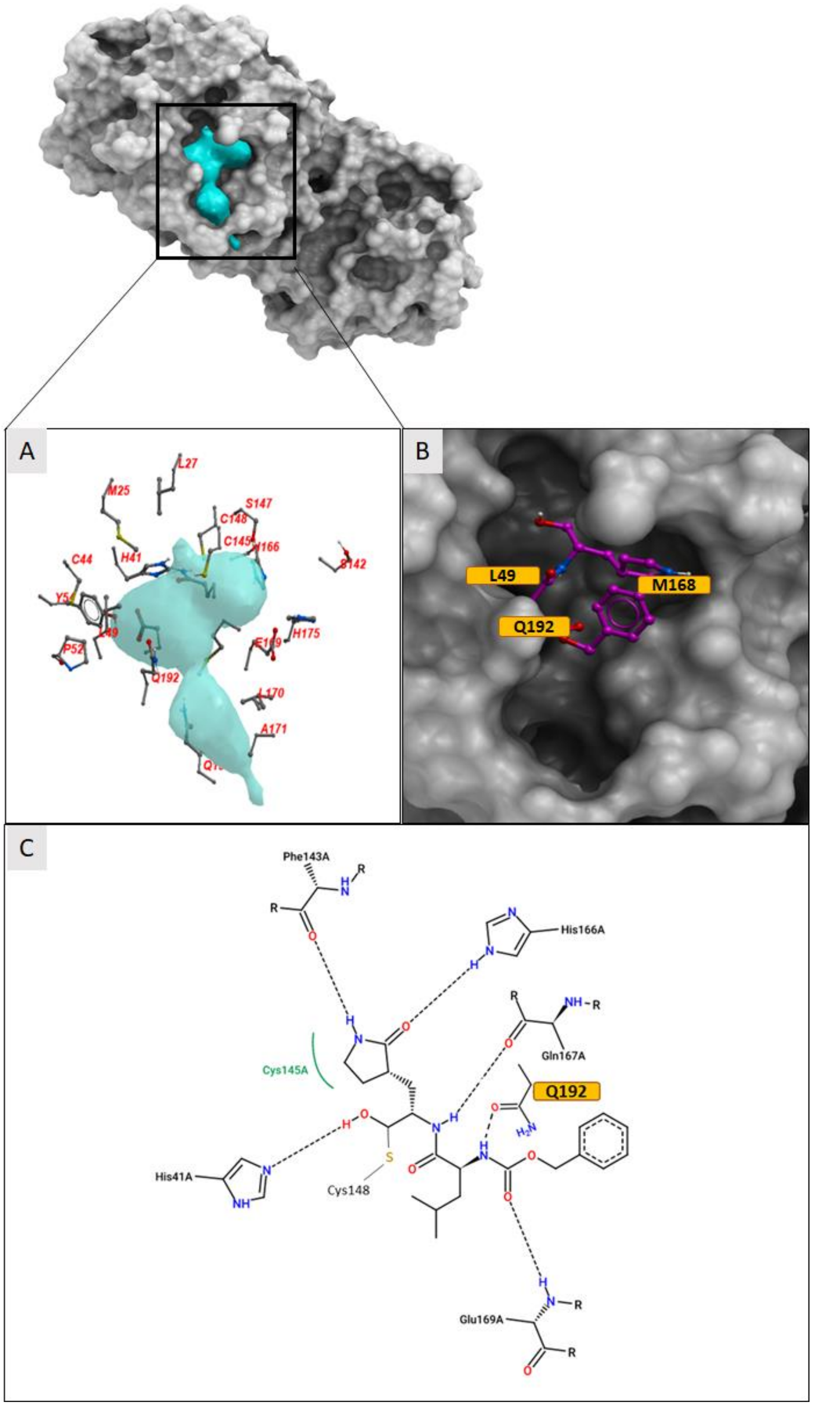


Figure 6. A. The catalytic site of MERS main protease (pdb code 5wkj). **B.** Non-conserved critical sidechains are highlighted in orange (residues 168,192 in MERS are residues 165,189 in SARS). Crystallized inhibitor: purple. **C.** 2D interaction diagram of a MERS main protease inhibitor (created using Proteins.plus PoseView, PDB code 5wkj). A non-conserved critical sidechain is highlighted in orange (residue 192 in MERS is residue 189 in SARS).

Step 5: Downloading the coronavirus sequences from the online database: https://bigd.big.ac.cn/gwh/browse/virus/coronaviridae on April 9, 2020.

We downloaded all available coronavirus sequences under Viruses; Riboviria; Nidovirales; Cornidovirineae; Coronaviridae on April 9, 2020.

Step 6: Creating pools of MPro sequences for the SARS-CoV-2 samples from patients, and for the coronaviruses MPro sequence.

We first extracted the sequence of the SARS-CoV-2 MPro crystal structure in ICM. We used this sequence as a reference to search for SARS-CoV-2 MPro sequences from the database we downloaded. The output was a list of sequences with a range of identity percentages (Seq ID) to the SARS-CoV-2 MPro sequence. We filtered 682 SARS-CoV-2 samples from COVID-19 patients by selecting all sequences that had Seq ID>98% relevant to our reference sequence. This high sequence identity cut-off was necessary, as SARS-CoV-1 sequences are 96% identical to SARS-CoV-2. Afterwards, we created a pool of samples of coronavirus sequences with Seq ID>40%. 4903 samples met this cutoff. We used this sequence identity cut-off to avoid including the other coronavirus (Papain-like) protease in this MPro-focused set. The schematic in Figure 7 depicts this concept visually.

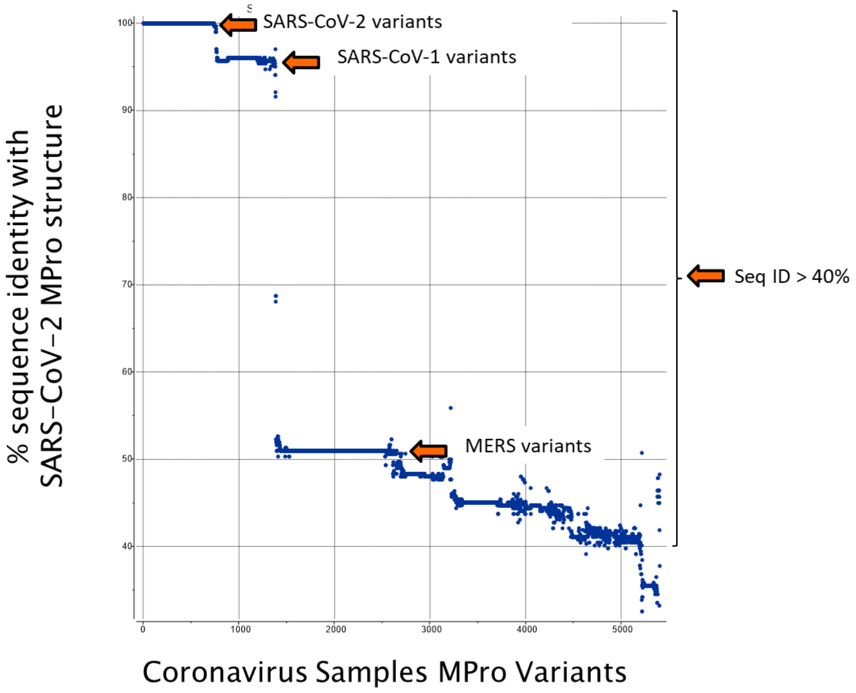
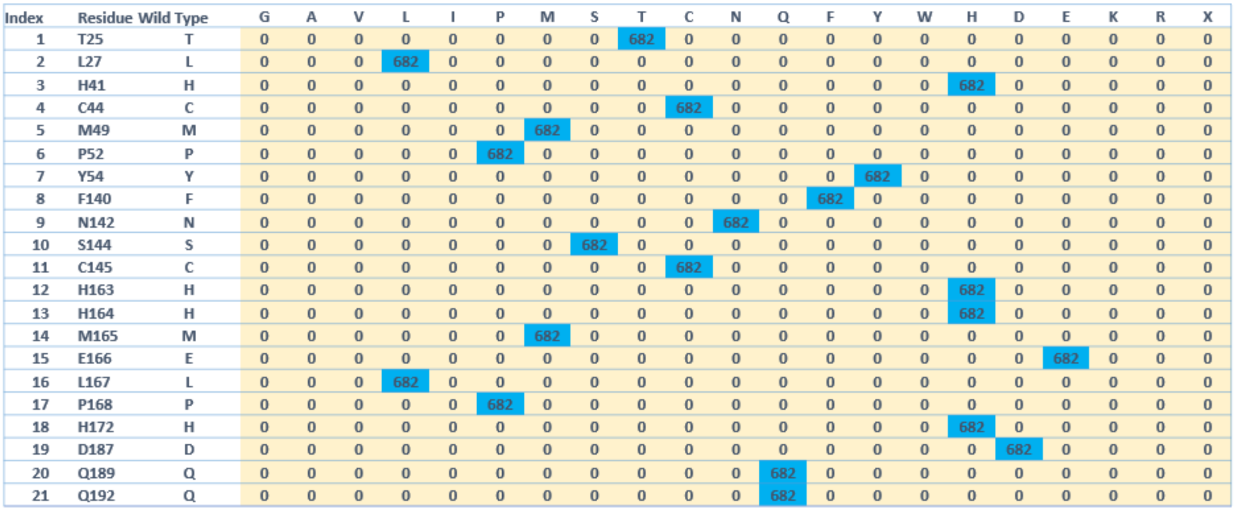


Figure 7. The Y-axis shows the % sequence identity (Seq ID) values, and the X-axis represents the coronavirus sample entries whose sequence were compared against the SARS-CoV-2 MPro sequence. All samples that met the cutoff of Seq ID>98% were grouped as SARS-CoV-2 samples. All samples with Seq ID>40% were grouped as coronavirus sequences for the MPro.

Step 7: Sequence diversity at the main protease catalytic site.

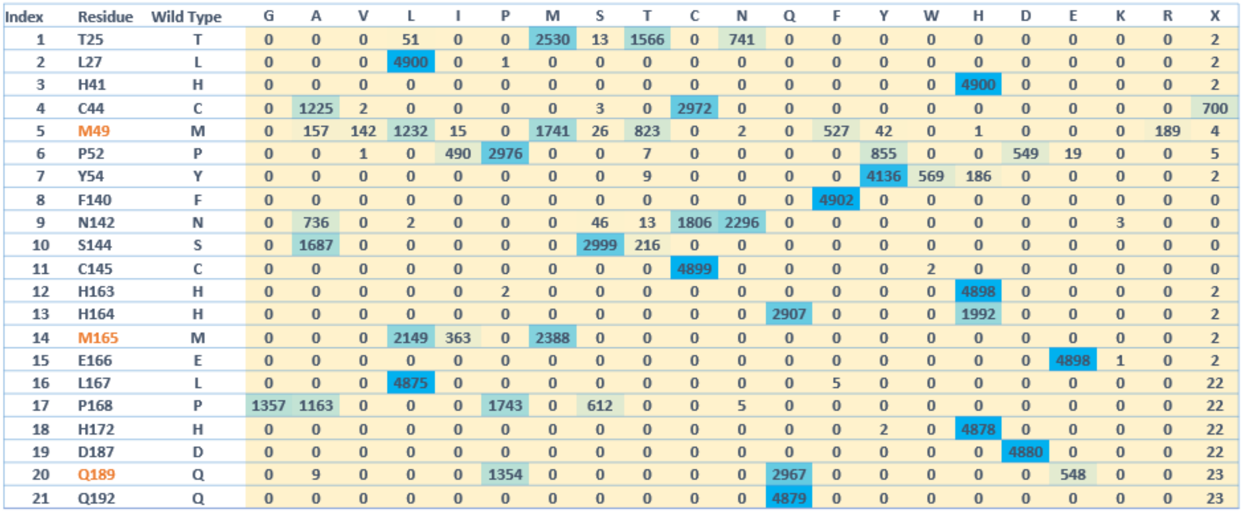
We next looked at the sequence diversity at the MPro catalytic site (21 residues) across 682 SARS-CoV-2 samples from COVID-19 patients which we selected in step 5. A matrix table was constructed to report the sequence diversity across these 21 residues. The result of this analysis is shown in Table 1. All 682 samples are identical at the main catalytic site and we saw no mutation.



Index	Residue	Wild Type	G	A	V	L	I	P	M	S	T	C	N	Q	F	Y	W	H	D	E	K	R	X
1	T25	T	0	0	0	0	0	0	0	0	682	0	0	0	0	0	0	0	0	0	0	0	0
2	L27	L	0	0	0	682	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	H41	H	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	682	0	0	0	0	0
4	C44	C	0	0	0	0	0	0	0	0	0	682	0	0	0	0	0	0	0	0	0	0	0
5	M49	M	0	0	0	0	0	0	682	0	0	0	0	0	0	0	0	0	0	0	0	0	0
6	P52	P	0	0	0	0	0	682	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
7	Y54	Y	0	0	0	0	0	0	0	0	0	0	0	0	0	682	0	0	0	0	0	0	0
8	F140	F	0	0	0	0	0	0	0	0	0	0	0	0	682	0	0	0	0	0	0	0	0
9	N142	N	0	0	0	0	0	0	0	0	0	0	682	0	0	0	0	0	0	0	0	0	0
10	S144	S	0	0	0	0	0	0	0	682	0	0	0	0	0	0	0	0	0	0	0	0	0
11	C145	C	0	0	0	0	0	0	0	0	0	682	0	0	0	0	0	0	0	0	0	0	0
12	H163	H	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	682	0	0	0	0	0
13	H164	H	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	682	0	0	0	0	0
14	M165	M	0	0	0	0	0	0	682	0	0	0	0	0	0	0	0	0	0	0	0	0	0
15	E166	E	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	682	0	0	0
16	L167	L	0	0	0	682	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
17	P168	P	0	0	0	0	0	682	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
18	H172	H	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	682	0	0	0	0	0
19	D187	D	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	682	0	0	0	0
20	Q189	Q	0	0	0	0	0	0	0	0	0	0	0	682	0	0	0	0	0	0	0	0	0
21	Q192	Q	0	0	0	0	0	0	0	0	0	0	0	682	0	0	0	0	0	0	0	0	0

Table 1. Sequence diversity at the main protease catalytic site across 682 SARS-CoV-2 samples downloaded from https://bigd.big.ac.cn/gwh/browse/virus/coronaviridae on April 9, 2020. The column "X" represents deletions.

Looking beyond the SARS-CoV-2 samples, we extended our analysis to the 4903 sequences with greater than 40% sequence similarity to the SARS-CoV-2 MPro sequence. Table 2 shows the sequence diversity across the 21 critical residues of the MPro catalytic site. We saw a significant variation at the catalytic site of MPro across the 4903 coronavirus samples.



Index	Residue	Wild Type	G	A	V	L	I	P	M	S	T	C	N	Q	F	Y	W	H	D	E	K	R	X
1	T25	T	0	0	0	51	0	0	2530	13	1566	0	741	0	0	0	0	0	0	0	0	0	2
2	L27	L	0	0	0	4900	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2
3	H41	H	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4900	0	0	0	0	2
4	C44	C	0	1225	2	0	0	0	0	3	0	2972	0	0	0	0	0	0	0	0	0	0	700
5	M49	M	0	157	142	1232	15	0	1741	26	823	0	2	0	527	42	0	1	0	0	0	189	4
6	P52	P	0	0	1	0	490	2976	0	0	7	0	0	0	0	855	0	0	549	19	0	0	5
7	Y54	Y	0	0	0	0	0	0	0	0	9	0	0	0	0	4136	569	186	0	0	0	0	2
8	F140	F	0	0	0	0	0	0	0	0	0	0	0	0	4902	0	0	0	0	0	0	0	0
9	N142	N	0	736	0	2	0	0	0	46	13	1806	2296	0	0	0	0	0	0	0	3	0	0
10	S144	S	0	1687	0	0	0	0	0	2999	216	0	0	0	0	0	0	0	0	0	0	0	0
11	C145	C	0	0	0	0	0	0	0	0	0	4899	0	0	0	0	2	0	0	0	0	0	0
12	H163	H	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	4898	0	0	0	0	2
13	H164	H	0	0	0	0	0	0	0	0	0	0	0	2907	0	0	0	1992	0	0	0	0	2
14	M165	M	0	0	0	2149	363	0	2388	0	0	0	0	0	0	0	0	0	0	0	0	0	2
15	E166	E	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4898	1	0	2
16	L167	L	0	0	0	4875	0	0	0	0	0	0	0	0	5	0	0	0	0	0	0	0	22
17	P168	P	1357	1163	0	0	0	1743	0	612	0	0	5	0	0	0	0	0	0	0	0	0	22
18	H172	H	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	4878	0	0	0	0	22
19	D187	D	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4880	0	0	0	22
20	Q189	Q	0	9	0	0	0	1354	0	0	0	0	0	2967	0	0	0	0	0	548	0	0	23
21	Q192	Q	0	0	0	0	0	0	0	0	0	0	0	4879	0	0	0	0	0	0	0	0	23

Table 2. Sequence diversity at the main protease catalytic site across 4903 coronavirus sequence samples downloaded from https://bigd.big.ac.cn/gwh/browse/virus/coronaviridae on April 9, 2020. The column "X" represents deletions. non-conserved residues critical for inhibitor binding are highlighted in orange.