Published in final edited form as:
Biochemistry. 2023 February 07; 62(3): 747–758. doi:10.1021/acs.biochem.2c00479.

Mutation Effects on Structure and Dynamics: Adaptive Evolution of the SARS-CoV-2 Main Protease

Elizabeth M. Diessner[†], **Gemma R. Takahashi**[‡], **Thomas J. Cross**[¶], **Rachel W. Martin**[§], **Carter T. Butts**[‖]

[†]Department of Chemistry, University of California, Irvine, Irvine, CA 92697, USA

[‡]Department of Molecular Biology & Biochemistry, University of California, Irvine, Irvine, CA 92697, USA

[¶]Department of Chemistry, University of California, Los Angeles, Los Angeles, CA 90095, USA

[§]Departments of Chemistry and Molecular Biology & Biochemistry, University of California, Irvine, Irvine, CA 92697, USA

[‖]Departments of Sociology, Statistics, Computer Science, and EECS, University of California, Irvine, Irvine, CA 92697, USA

Abstract

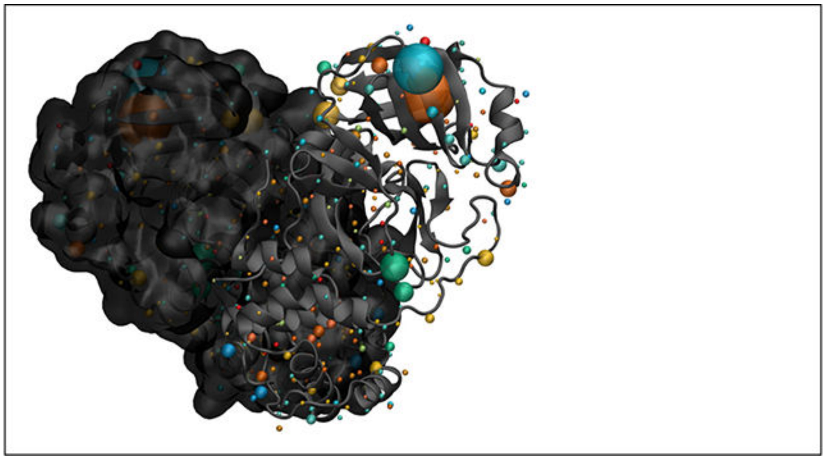
The main protease of SARS-CoV-2 (M^{pro}) plays a critical role in viral replication; although it is relatively conserved, M^{pro} has nevertheless evolved over the course of the COVID-19 pandemic. Here, we examine phenotypic changes in clinically observed variants of M^{pro}, relative to the originally reported wild-type (WT) enzyme. Using atomistic molecular dynamics simulations, we examine effects of mutation on protein structure and dynamics. In addition to basic structural properties such as variation in surface area and torsion angles, we use protein structure networks (PSNs) and active site networks (ASNs) to evaluate functionally relevant characters related to global cohesion and active site constraint. Substitution analysis shows a continuing trend toward more hydrophobic residues that is dependent on the location of the residue in primary, secondary, tertiary, and quaternary structure. Phylogenetic analysis provides additional evidence for the impact of selective pressure on mutation of M^{pro}. Overall, these analyses suggest evolutionary adaptation of M^{pro} toward more hydrophobicity and a less-constrained active site in response to the selective pressures of a novel host environment.

Graphical Abstract:

rwmartin@uci.edu; buttsc@uci.edu, Phone: +1 (949) 824-8591.

Supporting Information

Supplemental tables and figures referenced in the main text. Density plots comparing cohesion scores within each chain. Density plots comparing cohesion scores between chains, within each domain. Log ratio of torsion angle variance for the monomer and dimer conformations. Mean differences in side-chain properties of substituted residues. Distribution of conserved residues by amino acid type.



Introduction

The SARS-CoV-2 main protease (M^{pro}), also referred to as non-structural protein 5 (nsp5) or 3-chymotrypsin-like cysteine protease ($3CL^{pro}$), is a vital component of the coronavirus replication machinery[1]. During replication, the host ribosomes translate the SARS-CoV-2 non-structural proteins (nsps, i.e., enzymes) as a long polyprotein; this must then be cleaved into individual proteins to complete the expression and maturation process. In SARS-CoV and SARS-CoV-2, this cleavage function is performed by two proteases: the papain-like protease (PL^{pro}), and M^{pro} [2,3]. The first three cleavage sites, corresponding to the release of nsp1-nsp3, are cleaved by PL^{pro}, with the remaining 11 cleavage sites handled by M^{pro}, including those needed to release M^{pro} itself[4,5]. M^{pro} is thus necessary for maturation of the bulk of the proteins comprising the SARS-CoV-2 replicase[6]. M^{pro} also targets several proteins in the host cell, including key components of the cytokine and inflammatory responses[6,7].

M^{pro} itself is a cysteine protease, in which hydrolysis is performed by a catalytic dyad composed of a neutral (protonated) cysteine (C145) and a histidine (H41); this mechanism is strongly conserved among coronaviruses[1,8,9]. M^{pro}'s active conformation is a homodimer[1], although limited activity of M^{pro} monomers has been reported[10]. Despite its greatly reduced activity, molecular modeling suggests that the monomer is likely to be stable under physiological conditions, with a conformation that is similar to its conformation in the active homodimer[11]. Monomer and dimer structures, labeled by domain, are shown in Figure 1.

SARS-CoV-2 is believed to have transferred to the human population from zoonotic origin[13,14], and shares particular similarity with a number of bat coronaviruses[15]. While mutations to the infamous spike protein capture the attention of the public[16], other coronavirus proteins are also subject to evolutionary change, either due to neutral drift or as an adaptive response to environmental pressure. When adapting to a new host organism, selection pressure may be imposed by differences in the internal environment of host cells. For instance, bats experience a larger range of body temperatures compared with humans[17,18], including periods of activity at very high temperature[19,20]. Differences in host body temperatures impose different thermodynamic and kinetic constraints on the structure and activity of viral proteins within cells, which is a known factor limiting inter-species virus transmission[21,22] as well as tissue tropism within a single host[23,24].

As shown in studies of extremophilic organisms, the stability and catalytic efficiency of enzymes is dependent on their thermal environments[25,26]. Proteins in organisms that regularly experience high temperatures require stronger and more extensive interactions among residues, such as disulfide bonds and salt bridges to maintain stability[27–29], whereas proteins in low-temperature regimes require greater internal flexibility to facilitate catalysis[30]. The large and abrupt fluctuations in body temperature of bats are representative of frequent thermodynamic changes that put different kinds of stress on proteins, which may require particular structural responses to maintain structure and activity[30].

Beyond structural effects, mutations may also affect dynamics. Changes to local structure near the active site are particularly relevant, since such changes can affect both protein-substrate interactions and catalysis. Stronger side-chain interactions within the active site, for instance, may increase constraint on the dynamics of the catalytic residues. At the same time, long-range effects of residue substitution are known[31,32], suggesting that functionally relevant mutations may occur throughout the protein, as already observed for HIV protease[33] and SARS-CoV M^{pro} [34.]

For SARS-CoV-2 M^{pro}, then, selection for successful replication in a novel host environment is likely to favor systematic changes in protein structure and dynamics, which in turn will favor specific patterns of substitution. Such patterns may or may not be evident from sequence alone, because many different mutations may lead to similar physical properties; however, if present, selection pressure should manifest as consistent differences between structural and dynamic properties of WT M^{pro} versus ecologically successful mutants. By contrast, functionally critical properties that must be conserved between human and prior hosts would be expected to remain similar for both WT and successful variants, and properties under neutral drift would be expected to show variation with no systematic change from WT. Examination of structure and dynamics across a large range of ecologically successful mutants compared to WT thus provides evidence regarding adaptation by M^{pro} to its new environment. Early studies have suggested that some M^{pro} variants do differ from WT in structure and dynamics[11,35,36], motivating a systematic comparative analysis.

In this study, we identify evidence of selective pressure on the evolutionary adaptation of M^{pro} by analyzing results from molecular dynamics simulations and network analysis of all 1253 clinically identified variants of M^{pro} that were reported to the GISAID database over the first year of the COVID-19 pandemic (i.e., before February 25, 2021). Focusing on clinically observed variants allows us to work with mutations that were both functional and ecologically successful, in that they could successfully infect human hosts "in the wild." To distinguish between effects arising directly from changes to the structure of the M^{pro} monomer and those emerging only in the dimeric state, we examine models of both the functional dimer and the free monomer in solution. Trends in physical properties of variants relative to WT are assessed using multiple techniques. Relative Solvent Accessibility (RSA) is used to calculate total surface area, providing preliminary information on the effect of mutation on global structure. Internal changes to structure are further investigated by analysis of Protein Structure Networks (PSNs) to observe changes in internal residue interaction rates. The effects of substitutions on local dynamics are observed by comparing variation in torsion angles - extracted from dynamic simulation trajectories - between and within variants. Active Site Networks (ASNs) of each variant are constructed to measure local constraint on the active site. Finally, we investigate trends in the physical properties of amino acid substitutions, and explore the ways the location of certain substitutions - or lack thereof - contribute to a response to selective pressure that may be guiding the adaptive evolution of M^{pro}.

The results of the following analyses provide a rich context for understanding the physical adaptation of M^{pro}, and suggest a number of targets for experimental investigation, which will be required to probe the impact of the observed mutations on catalytic activity and kinetic parameters. Compared with WT, variants are observed on average to have more solvent-accessible surface area (SASA), indicating either an increase in size of surface residues or a loosening of internal structure. In the monomeric state, M^{pro} is observed to have lower cohesion overall, contributing to the loosening of the structure, while the dimeric state conserves internal interactions in domain 2. Backbone torsion angles are generally similar between the monomeric and dimeric states, with mutations having the greatest impact on the backbone structure of residues in domain 2 of both states. The two active sites of the dimeric state trend towards less constraint on the catalytic residues, but the monomeric state shows no definite trend, despite the similar effects of mutation on the structure of the monomeric and dimeric states. The substitutions themselves generally trend towards more hydrophobic residues, with certain frequently occurring mutations near the active site showing a trend toward more hydrophilic residues. Frequently observed mutations, including some located near the active site, have occurred in several unique branches, indicating a possible benefit to M^{pro} function that is supported by selective pressure on the enzyme.

Methods

Sequence Preprocessing

Human-derived SARS-CoV-2 full genome sequences were retrieved from the GISAID EpiCoV database[37] on February 25, 2021 at 10:15 AM (PST). These were filtered for size and quality; those with <1 percent N content and lengths within +/−3 percent of the length of a designated WT sequence (RefSeq: NC 045512.2[13]) (29,006 bp–30,800 bp inclusive) were retained for further processing. High-quality sequences were filtered for valid M^{pro} sequences, and then again for modellable M^{pro} sequences. For our purposes, “valid” sequences refer to those with no frameshifts, deletions, insertions, Ns, or non-standard IUPAC nucleotides (those other than A, C, U, G); “modellable” sequences are valid M^{pro} sequences with no non-synonymous mutations that result in either changes to the active site (H41 or C145) or premature stop codons, as the true functionality and/or translated structures of these variants are currently unknown. These M^{pro} sequences were located in and extracted from full genomes by using six 15-nucleotide keys, derived from the

NC_045512.2 reference M^{pro} sequence (loc: 10,055–10,972). All sequence preprocessing was done using custom scripts in Python (v3.7.6)[38].

Alignments

All full genome alignments were performed using suggested MAFFT (v7.471)[39] protocols for SARS-CoV-2 (https://mafft.cbrc.jp/alignment/software/closelyrelatedviralgenomes.html). Full genomes were aligned to a WT reference (NC_045512.2), using the options "–auto" and "addfragments"; in order to retain site information for phylogenetic analysis, the "–keeplength" option was not used.

Clustering and Phylogenetic Tree

A phylogenetic tree was constructed for aligned full genomes that contained non-WT, modellable M^{pro} variants using FastTree (v2.1.11 SSE3)[40,41] with OpenMP[42] (FastTreeMP); the "-fastest" option was used. This included 70,246 full genomes with non-synonymous M^{pro} mutations (considered "variants") and 34,909 full genomes with synonymous M^{pro} mutations (same protein sequence as WT). One WT full genome reference, (NC_045512.2) was also included. Visualizations were generated in R (v4.0.4)[43] using ggtree[44], ape[45], ggplot2[46], treeio[47], tidyverse[48], ggtreeExtra[44], aplot[49], data.table[50], svglite[51].

Molecular Modelling of WT and Variant Structures

Monomer and dimer conformations of variant structures were predicted with MODELLER 9.23[52] using the PDB structure 6Y2E[12] as the WT template. All structures underwent three rounds of annealing and MD refinement using "slow" optimization. The protonation states were corrected for the predicted cell environment using PROPKA 3.1[53]. The corrected structures were minimized and equilibriated in explicit solvent. MD trajectories were then simulated from the corrected structures using NAMD[54] with a CHARMM36[55] force field and TIP3P water at 310 K under periodic boundary conditions for a water box with a 10 Å margin in an NpT ensemble. Solvated models were energy-minimized for 10,000 iterations, then simulated once for 10 ps to make water box size adjustments (for PME calculations), and once more for a 10ns trajectory with sampled conformations saved every 20 ps. Temperature control was maintained via Langevin dynamics with a damping coefficient of 1/ps, and pressure control was performed via a Nose-Hoover Langevin[56] piston set at 1 atm. Visualizations and other static analyses are based on the final conformations from each trajectory, with full trajectories used for dynamic analyses. Visualizations were performed using VMD[57]. Solvent accessible surface area calculations were performed using the `dssp.pdb` function in the bio3d library in R[58].

Network Analysis

All frames from each respective simulated trajectory were individually translated into PSNs using scripts written using the statnet, Rpdb, and bio3d libraries in R[58–61]. Vertices for each network follow the convention established by Benson and Daggett [62] - atoms are grouped into chemical moieties, each of which is represented by a node. Each residue is thus represented by a collection of nodes, and an edge (tie) is formed between two nodes when there exist respective atoms associated with each node that lie within a threshold distance of each other in the selected frame. The distance cutoff used here is 1.1 times the sum of the respective van der Waals radii of the two atoms. An ASN[63] was constructed for the active site of each variant structure by inducing a subgraph comprised of the nodes representing Cys 145, His 41, and all adjacent vertices from the respective PSN. PSNs and ASNs were calculated for all frames from each trajectory, all of which were used in the reported analyses.

Analyses of the PSNs used degree k-cores[64] to characterize the cohesion of each monomer and dimer chain, with the core number of each node (i.e., the highest k such that the node belongs to the kth core) being employed as a measure of local cohesion. Mean core numbers for vertices within each domain, and for the protein as a whole, were used to assess cohesion; all quantities were computed within each frame, with trajectory averages used as for structural comparison. Autocorrelation-corrected bootstrap standard errors were calculated to control for within-trajectory temporal autocorrelation in the trajectory means, and variant values were treated as significantly different from WT if they differed by more than two standard errors. Calculations were done using the sna library in R[65]. Analyses of ASNs included calculations of degree, triangle degree, core number, and connectivity, each averaged over the active site. Here, degree refers to the number of ties a particular node has - i.e. the total number of contacts. Triangle degree refers to the number of triangles containing a particular vertex, and core number for these analyses was assessed within-ASN (as opposed to core number within the broader PSN). Connectivity was measured using the log of the number of indirect paths between the two active site residues. Together, degree, triangle degree, core number, and connectivity give an indication of the freedom of movement within the active site. This gives an approximation of an active site state, which can be used to distinguish active site conformations which are more “open” or “closed.” Quantitatively, we assess this via a *constraint score,* which is the score of each network on the first principal component of the combined and standardized degree, triangle degree, core number, and connectivity measures.

Results and Discussion

Variants Tend Toward Less Compact M^{pro} Structure

Surface area increases, but more so in the monomer than the dimer.—Overall, the most common effect of mutations on the monomer conformation is to increase the surface area of the enzyme, as shown in Figure 2, with 46.3% of variants with increased surface area (p-value = 0.01 using an exact binomial test), 53.1% with no change (p-value = 0.03), and 0.6% with decreased surface area (p-value $<2.2\times10^{-16}$). This could be a side effect of bulkier residues, or the result of a decrease in internal interactions. Alternatively, bulky and hydrophobic residue substitutions in the interior could cause the structure to expand outward to accommodate the larger side-chains.

The increase in surface area of the monomer is less pronounced in the dimeric conformation: although we do see a net tendency towards SASA increase (28.4% increased, 7.4% decreased, and 64.6% stay the same, p-values $<2.2\times10^{-16}$ using an exact binomial test), fewer variants show significant differences, and the location of WT within the distribution is less skewed. This suggests that surface enlargement occurs disproportionately within the dimerization interface, resulting in a total surface area that is more conserved upon dimerization. That said, we still observe a significant bias towards higher-SASA dimers, which is consistent with selection favoring a somewhat looser, enlarged protein surface.

Global cohesion is lower in the majority of variants, except for domain 2 in the dimeric state.—Looking at the impact of substitution on cohesion within free M*pro* monomers, we see a consistent pattern of structural "loosening" relative to WT, with 78.5% of variants showing significantly lower levels of cohesion, versus 0.3% showing higher levels (p-value $<2\times10^{-16}$ using an exact binomial test). PSNs measuring internal interaction rates between moieties show a decrease in internal cohesion in all domains of the monomer (Fig. 3), with a slightly reduced degree of loosening in domain 2. This suggests selection for increased flexibility at the level of individual proteins, possibly as a result of the more moderate thermal environment of the human host.

Is this monomer-level change retained upon dimerization? Fig. 4 shows that this pattern of reduced cohesion is largely preserved, with looser structures seen in entire dimerized chains, as well as internally within domains 1 and 3. Domain 2, however, shows a rather different pattern, with no clear evolutionary trend: indeed, a substantial fraction (33.1% in the high-cohesion chain and 13.6% in the low-cohesion chain, p-values $<2\times10^{-16}$ using an exact binomial test) actually show enhanced cohesion versus WT. The presence of diversification (with some variants higher, others lower, and relatively few remaining similar) is compatible with the notion that domain 2 within the dimeric state is not being actively selected with respect to cohesion, and is subject to neutral drift. It is interesting to observe in this regard that we do see a cohesion-reducing trend for domain 2 in the monomer, and thus that the apparent direction of evolution is different for the components of active M*pro* versus the active dimeric state itself; one plausible explanation is that the monomeric loosening within domain 2 arises as a side effect of overall selection for a less cohesive protein, but that interactions in the dimer interface do not preserve this property for that region in the dimeric state. Either way, we find no evidence that M*pro* is being selected for a looser domain 2 structure in the dimer.

Local structural changes due to mutations show similar effects for free and dimerized monomers, despite cohesion differences.—To assess local changes in backbone structure due to residue substitution, we compute the (angular) mean and variance for each backbone torsion angle in each trajectory for both free monomers and dimers. Using this, we compare the variance in angles within trajectory versus across trajectories, allowing us to determine the extent to which local structure differs across variants above and beyond natural variations due to protein dynamics. Figure 5 shows the log-ratio of the between-variant versus within-variant angular variance, plotted by residue. High log-ratio values (blue areas) show substantial sensitivity to mutations, while low log-ratio values (red areas) show little structural change relative to normal fluctuations due to protein dynamics. We see here that the bulk of the mutation effects are in or adjacent to domain 2, with domain 3 showing particularly low levels of sensitivity to observed substitutions. Taking these results in the context of the above findings regarding cohesion, we conclude that the cohesion changes seen in domains 1 and 3 are not due primarily to local deformation of the backbone in these regions of the protein, but more plausibly to a combination of side chain interactions and interactions with domain 2 residues (which do show greater change in torsion angle). Local deformation in domain 2 may thus be less important for the impact it has on domain 2 itself (which, as seen above, is inconsistent), versus its effect on the network of contacts in the neighboring domains (which both show consistent patterns of change).

Figure 5 also reveals that the pattern of backbone structure change in the free monomer is extremely similar to what is observed in the dimerized monomer, indicating that local structural changes are not strongly affected by dimerization. The immediate impact of mutation on local (backbone) structure thus depends only on interactions that are internal to the M^{pro} monomer itself, and are not related to interactions across the dimerization interface.

Mutations Increase Active Site Flexibility in the Active Dimer State

Mutations increase active site flexibility in the dimer, but not the free monomer.—If mutations were selected to increase function of free monomers, the local structure around the active site of the monomer would be expected to show systematic change. This is not the case. As shown in Figure 6, constraint on the active site of the monomer does not trend in any direction; the presence of a large number of variants with either significantly higher (23.7%) or lower (11.2%) constraint levels suggests drift rather than conservation (p-values $<2.2\times10^{-16}$ using an exact binomial test). By contrast, we see evidence of systematic selection for lower levels of active site constraint (looser structure) in the dimeric state. Not only are the grand means across variants lower for dimer active sites, but the majority (59.7% in higher scoring chain, 69.1% in lower scoring chain, p-values $<7.5\times10^{-12}$, $<2.2\times10^{-16}$, respectively using an exact binomial test) of variants have mean constraint scores that are significantly below WT. The presence of large differences in the dimer vs. monomer sites indicates that active site loosening is not driven by local structural changes to the monomer itself, but instead emerges from interaction between monomers in the dimer.

The decrease in constraint of the dimer active sites supports the hypothesis that the enzyme is increasing flexibility to adapt to the cellular environment of the human host. The difference between the changes in the properties of the dimer versus free monomer sites further sheds light on the dramatically higher activity of M^{pro} in the dimeric state: although earlier work[11] has shown that monomeric active site conformations do not differ markedly from dimeric ones, dimerization clearly shifts the equilibrium distribution of conformational states. Selection in this case appears to be operating on this shift, rather than on the underlying distribution, resulting in a pattern of changes that is selective for dimers while apparently neutral for free monomers.

Amino Acid Substitutions Favor Increased Size and Hydrophobicity

In general, substitutions increase hydrophobicity.—Out of the 306 residues of the mature M^{pro} sequence, 269 have been substituted in at least one variant. To analyze the trends in properties of the substituted amino acids a substitution network was created by forming an adjacency matrix of substitutions. The rows and columns of the matrix were labeled with the 20 unique amino acids, and values in the matrix represented the frequency of each substitution occurring in the set of 1253 variants. This resulted in the network shown in Figure 7. Nodes represent unique amino acids, and edges represent the frequency of respective substitution. Substituted amino acids tend to be more hydrophobic and massive than their predecessors.

The large number of substitutions between certain residues, such as $L \rightarrow F$, $K \rightarrow R$, $G \rightarrow S$, and $A \rightarrow V$ indicate that these substitutions are highly favorable. These four substitutions are all examples of an exchange for a bulkier residue, and in the case of $G \rightarrow S$, a *more hydrophilic* residue. In the cases of $L \rightarrow F$ and $K \rightarrow R$, the substituted residues are able to form more complex intermolecular interactions, with a wider range of pi-stacking and cation-pi interactions available compared to the starting residues.

Frequent substitutions near the active site are either similar in hydrophobicity or more hydrophilic, while those in domain 2 are more hydrophobic.—Figure 8 shows the frequencies of variants containing a substitution at a particular residue. The three most common substitutions among variants, L89F, K90R, and G15S, all occur in domain 1, and are all substitutions for bulkier residues. The decrease in cohesion of domain 1 could be caused by an increase in solvent interactions due, in part, to these three substitutions.

The first rug in Fig. 8 is the mean change in hydrophobicity, and shows a greater occurrence of hydrophobic substitutions in domain 2 than in either other domain. This also coincides with residues being more buried, as shown in the second rug by the darker blue coloring. An increase in the hydrophobicity of buried residues in domain 2 could be a response to a decreasing hydrophobic effect required to maintain the cohesion needed for certain dynamics resulting from internal interactions occurring between the dimer interface and the active site. While there are some structural changes upon dimerization in domain 2 due to substitutions, as seen in Fig. 5, those substitutions tend to be for more hydrophobic residues that are participating in the dimer interface. Increasing hydrophobicity at the dimer interface would result in increased contact between the two chains due to the hydrophobic effect. Additionally, the location of more hydrophilic substitutions in regions where the chain is transitioning from the interior to the surface would cause a decrease in cohesion as those residues have stronger solvent interactions. Such regions are found in all three domains.

Persistent substitutions - those that occur most frequently - are shown in their location on one chain of the dimer in Figure 9. The most frequent substitution, L89F, is located between the folded β-sheets of domain 1. The substitution with a bulkier residue, phenylalanine, would push the β-sheets apart, reducing the cohesion of domain 1 and pulling the catalytic His41 back towards the β-fold. This change in structure of domain 1 would affect interactions between residues 43–50 and residues 186–190 on the unstructured loop between domains 2 and 3. There may also be some effect on the domain 1 residues near the N-terminus.

The substitution K90R would be expected behave similarly to L89F. However, this residue is facing out from the surface of the protein, and the substitution to arginine from lysine increases the number of potential hydrogen bonds that can be formed with the solvent in addition to increasing the bulk of the side chain. This may cause the domain 1 β-fold to be pulled open from the outside, instead of pushed from the inside. Variants with the K90R substitutions may see less effect on the interactions between residues 43–50 and 186–190, and more impact on the cohesion of domain 2 due to interactions between residues 97–105 in the unstructured loop between domains 1 and 2. The substitutions of G15S and G71S occur much closer to the dimer interface. Glycine and serine are both highly flexible residues, so the substitution at these locations may not have an appreciable effect on local structure. However, the polar nature of serine may cause it to respond to dynamics of other residues. For instance, a serine at residue 15 or 71 may interact with the polar hydroxyl group on Y154 of the opposite dimer chain, which would cause some correlation between the dynamics of domain 1 of one chain and domain 2 of the opposite chain.

P108 and P132 together form the ends of a loop that extends through domain 2 to interact at the dimer interface, forming a large part of the dimer interface. The location of the P108S and P132S substitutions may optimize their effect due to their connections with the dimer interface and proximity to the active site. Persistent mutations that are more hydrophilic are located away from the dimer interface, or else function to maintain the location of the interface by becoming less susceptible to the hydrophobic effect. These mutations are all located in domain 1, yet have limited impact on the active site except when in the dimer conformation. The more hydrophobic of the persistent mutations are located in domain 2, and have an influence on interaction at the dimer interface, while also having limited impact on the active site. The increasing hydrophobicity due to substitutions in domain 2 contributes to the conserved cohesion of the domain, as well as the increased influence of the dimer interface on local structure.

Conserved residues are concentrated in domain 2, and tend to be polar.— Substitutions in domain 2 for more hydrophobic residues may help to maintain the cohesion of the structure, as well as the dynamics resulting from interactions between the dimer interface and active site. Conserved residues may facilitate those dynamics to such an extent that any substitution that disrupts those interactions would inhibit function of the protein. This hypothesis is supported by the pattern of conserved residues in the dimer structure, shown in Figure 10.

Conserved residues in domain 2, located between the dimer interface and the active site, tend to be aromatic polar and nonpolar residues. Nonpolar residues that are conserved in domains 1 and 3 are by contrast non-aromatic. Acidic residues that are conserved are concentrated at the dimer interface near the N-termini, as well as on domain 2 at the active site. Polar residues other than Gly are concentrated around the active site, and at the dimer interface near the C-termini.

Relationships between Variants

Clustering in phylogenetic tree shows independent occurrence of frequent mutations, supporting the selective pressure hypothesis.—Frequent mutation is a form of adaptation in viruses[66], but while many rare variants exist in the population through luck, those that are observed in large numbers may be evidence of selective pressure[67].

Clustering patterns (Figure 11) have shown several large groups of recurring variants across disconnected lineages, supporting the hypothesis that this variation in sequence space may have also led to functional differences.

The most numerous mutations within the sample are, in increasing order: G71S, G15S, K90R, P108S, and L89F. These five were all present in a previous dataset from April, 2020[11], though their prevalence in certain SARS-CoV-2 lineages were not necessarily as pronounced. Notably, G15S and K90R, which once dominated datasets over one year ago, have since been overtaken by L89F. Despite differences in raw counts, all five of these long-established mutations inhabit their own evolutionarily distinct clusters within the phylogenetic tree, often mimicking the large subtrees we saw in April, 2020 that were indicative of separate evolutionary events. Additionally, highly prolific mutations, including these five, have continued to remain viable in the population, co-occurring with secondary non-synonymous mutations that may impart their own structural or functional differences. For example, there are now 202 unique L89F variants (201 with at least one other amino acid mutation); in terms of mutational space, this means that nearly 1/6 of our unique variant dataset contains an L89F mutation. Although there is some overlap with other prominent mutations, much of that space is also taken up by G71S (36 variants), P108S (46 variants), G15S (74 variants), and K90R (97 variants).

Whole genome phylogenies are a useful tool in the study of viral evolution, but phylogenetic inferences should be made with the understanding that complex evolutionary dynamics are inherently difficult to capture. While neutral drift and selective mechanisms vie for control of genotypic diversity[68], factors like sequencing errors and sampling bias can disrupt attempts to accurately quantify their effects[69,70]. The study of SARS-CoV-2 in particular is further complicated by large numbers of sequences with low sequence variation[70], making it difficult to draw meaningful conclusions from phylogenetic analyses alone. Because of regional variation in sequencing rates and pandemic policy, it is difficult to know if the rise of certain variants is truly due to fitness, as is often suspected. However, the trends observed here in total surface area, cohesion, torsion angle variance, and active site constraint speak to adaptations resulting from selective pressure, and reinforce evidence to that end observed in M^{pro}'s phylogeny.

Conclusion

Taken together, our analyses suggest that the SARS-CoV-2 main protease is evolving in response to selective pressure, possibly brought by the difference in cellular environments of bats and humans. The resulting adaptations are observed to affect the global structure and active site dynamics of the dimer conformation differently than the free monomer, despite having similar impacts on the local backbone structure of both states; in the case of active site constraint, the observed pattern of change vs. wild type is seen only in the dimeric state, and thus emerges from interactions between monomers. Adaptations tend to conserve interactions at the dimer interface and in domain 2, while allowing the rest of the protein, including the active site, to become more flexible in the dimeric state. The locations and properties of frequently occurring substitutions, as well as that of conserved residues, help elucidate the relationship between structure, dynamics, and function of M^{pro} as it is revealed by the process of selective adaptation.

As with any computational study, a major function of this work is to suggest targets for experimental investigation. Our findings suggest both general trends to be tested, and variants predicted to have extremal properties (relative to the ensemble); both tests of these hypothesized trends and examination of the relationship between the structural characteristics considered here and catalytic function would both shed light on M^{pro} evolution and help guide future computational studies. We also note that a number of other nsps (including the papain-like protease, PL^{pro}[71]) are also highly conserved within the beta-coronaviruses,[72] suggesting mutation rates low enough to make computational studies like this one possible for such systems. Comparative analysis of changes seen across SARS-CoV-2 nsps in response to human host adaptation could provide deeper insights into ways in which evolutionary processes influence the molecular machines that carry out viral replication.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This research was supported by NSF awards 1633631 to E.M.D., and DMS-1361425 to C.T.B. and R.W.M., NASA Award 80NSSC20K0620 to R.W.M. and C.T.B., NIH grant 2R01EY021514 to R.W.M., the CIFAR Molecular Architecture of Life program, and the UC Irvine School of Physical Sciences Dean's Excellence Fund.

References

(1). Anand K; Palm GJ; Mesters JR; Siddell SG; Ziebuhr J; Hilgenfeld R Structure of Coronavirus Main Proteinase Reveals Combination of a Chymotrypsin Fold with an Extra α-Helical Domain. The EMBO Journal 2002, 21, 3213–3224. [PubMed: 12093723]

(2). Ziebuhr J In Coronavirus Replication and Reverse Genetics; Enjuanes L, Ed.; Current Topics in Microbiology and Immunology; Springer: Berlin, Heidelberg, 2005; pp 57–94.

(3). Song Z; Xu Y; Bao L; Zhang L; Yu P; Qu Y; Zhu H; Zhao W; Han Y; Qin C From SARS to MERS, Thrusting Coronaviruses into the Spotlight. Viruses 2019, 11, 59. [PubMed: 30646565]

(4). Qiu Y; Xu K Functional Studies of the Coronavirus Nonstructural Proteins. STEMedicine 2020, 1, e39–e39.

(5). Yan S; Wu G Potential 3-Chymotrypsin-like Cysteine Protease Cleavage Sites in the Coronavirus Polyproteins Pp1a and Pp1ab and Their Possible Relevance to COVID-19 Vaccine and Drug Development. The FASEB Journal 2021, 35, e21573. [PubMed: 33913206]

(6). Graham RL; Sparks JS; Eckerle LD; Sims AC; Denison MR SARS Coronavirus Replicase Proteins in Pathogenesis. Virus Research 2008, 133, 88–100. [PubMed: 17397959]

(7). Meyer B et al. Characterising Proteolysis during SARS-CoV-2 Infection Identifies Viral Cleavage Sites and Cellular Targets with Therapeutic Potential. Nature Communications 2021, 12, 5553.

(8). Anand K; Ziebuhr J; Wadhwani P; Mesters JR; Hilgenfeld R Coronavirus Main Proteinase (3CLpro) Structure: Basis for Design of Anti-SARS Drugs. Science 2003, 300, 1763–1767. [PubMed: 12746549]

(9). Wang F; Chen C; Tan W; Yang K; Yang H Structure of Main Protease from Human Coronavirus NL63: Insights for Wide Spectrum Anti-Coronavirus Drug Design. Scientific Reports 2016, 6, 22677. [PubMed: 26948040]

(10). Shi J; Wei Z; Song J Dissection Study on the Severe Acute Respiratory Syndrome 3C-like Protease Reveals the Critical Role of the Extra Domain in Dimerization of the Enzyme: DEFINING THE EXTRA DOMAIN AS A NEW TARGET FOR DESIGN OF HIGHLY SPECIFIC PROTEASE INHIBITORS *. Journal of Biological Chemistry 2004, 279, 24765–24773. [PubMed: 15037623]

(11). Cross TJ; Takahashi GR; Diessner EM; Crosby MG; Farahmand V; Zhuang S; Butts CT; Martin RW Sequence Characterization and Molecular Modeling of Clinically Relevant Variants of the SARS-CoV-2 Main Protease. Biochemistry 2020, 59, 3741–3756. [PubMed: 32931703]

(12). Zhang L; Lin D; Sun X; Curth U; Drosten C; Sauerhering L; Becker S; Rox K; Hilgenfeld R Crystal Structure of SARS-CoV-2 Main Protease Provides a Basis for Design of Improved α-Ketoamide Inhibitors. Science 2020, 368, 409–412. [PubMed: 32198291]

(13). Wu F et al. A New Coronavirus Associated with Human Respiratory Disease in China. Nature 2020, 579, 265–269. [PubMed: 32015508]

(14). Andersen KG; Rambaut A; Lipkin WI; Holmes EC; Garry RF The Proximal Origin of SARS-CoV-2. Nature Medicine 2020, 26, 450–452.

(15). Temmam S et al. Bat Coronaviruses Related to SARS-CoV-2 and Infectious for Human Cells. Nature 2022, 604, 330–336. [PubMed: 35172323]

(16). Callaway E Could New COVID Variants Undermine Vaccines? Labs Scramble to Find Out. Nature 2021, 589, 177–178. [PubMed: 33432212]

(17). Skåra KH; Bech C; Fjelldal MA; van der Kooij J; Sørås R; Stawski C Energetics of Whiskered Bats in Comparison to Other Bats of the Family Vespertilionidae. Biology Open 2021, 10, bio058640. [PubMed: 34338281]

(18). Ramos Pereira MJ; Stefanski Chaves T; Bobrowiec PE; Selbach Hofmann G How Aerial Insectivore Bats of Different Sizes Respond to Nightly Temperature Shifts. International Journal of Biometeorology 2022, 66, 601–612. [PubMed: 34817674]

(19). Maloney SK; Bronner GN; Buffenstein R Thermoregulation in the Angolan Free-Tailed Bat Mops Condylurus: A Small Mammal That Uses Hot Roosts. Physiological and Biochemical Zoology 1999, 72, 385–396. [PubMed: 10438676]

(20). Czenze ZJ; Naidoo S; Kotze A; McKechnie AE Bat Thermoregulation in the Heat: Limits to Evaporative Cooling Capacity in Three Southern African Bats. Journal of Thermal Biology 2020, 89, 102542. [PubMed: 32364970]

(21). Massin P; van der Werf S; Naffakh N Residue 627 of PB2 Is a Determinant of Cold Sensitivity in RNA Replication of Avian Influenza Viruses. Journal of Virology 2001, 75, 5398–5404. [PubMed: 11333924]

(22). Mänz B; Schwemmle M; Brunotte L Adaptation of Avian Influenza A Virus Polymerase in Mammals To Overcome the Host Species Barrier. Journal of Virology 2013, 87, 7200–7209. [PubMed: 23616660]

(23). V’kovski P et al. Disparate Temperature-Dependent Virus–Host Dynamics for SARS-CoV-2 and SARS-CoV in the Human Respiratory Epithelium. PLOS Biology 2021, 19, e3001158. [PubMed: 33780434]

(24). Shaw Stewart PD; Bach JL Temperature Dependent Viral Tropism: Understanding Viral Seasonality and Pathogenicity as Applied to the Avoidance and Treatment of Endemic Viral Respiratory Illnesses. Reviews in Medical Virology 2022, 32, e2241. [PubMed: 33942417]

(25). Boob M; Wang Y; Gruebele M Proteins: “Boil ‘Em, Mash ‘Em, Stick ‘Em in a Stew”. The Journal of Physical Chemistry B 2019, 123, 8341–8350. [PubMed: 31386813]

(26). Mathew GM; Madhavan A; Arun KB; Sindhu R; Binod P; Singhania RR; Sukumaran RK; Pandey A Thermophilic Chitinases: Structural, Functional and Engineering Attributes for Industrial Applications. Applied Biochemistry and Biotechnology 2021, 193, 142–164. [PubMed: 32827066]

(27). Kumar S; Nussinov R How Do Thermophilic Proteins Deal with Heat? Cellular and Molecular Life Sciences CMLS 2001, 58, 1216–1233. [PubMed: 11577980]

(28). Jorda J; Yeates TO Widespread Disulfide Bonding in Proteins from Thermophilic Archaea. Archaea 2011, 2011, e409156.

(29). González-Castro R; Gómez-Lim MA; Plisson F Cysteine-Rich Peptides: Hyperstablé Scaffolds for Protein Engineering. ChemBioChem 2021, 22, 961–973. [PubMed: 33095969]

(30). Suka A; Oki H; Kato Y; Kawahara K; Ohkubo T; Maruno T; Kobayashi Y; Fujii S; Wakai S; Lisdiana L; Sambongi Y Stability of Cytochromes c′ from Psychrophilic and Piezophilic Shewanella Species: Implications for Complex Multiple Adaptation to Low Temperature and High Hydrostatic Pressure. Extremophiles 2019, 23, 239–248. [PubMed: 30689055]

(31). Xie W; Nangle LA; Zhang W; Schimmel P; Yang X-L Long-Range Structural Effects of a Charcot–Marie–Tooth Disease-Causing Mutation in Human Glycyl-tRNA Synthetase. Proceedings of the National Academy of Sciences 2007, 104, 9976–9981.

(32). Axe JM; Boehr DD Long-Range Interactions in the Alpha Subunit of Tryptophan Synthase Help to Coordinate Ligand Binding, Catalysis, and Substrate Channeling. Journal of Molecular Biology 2013, 425, 1527–1545. [PubMed: 23376097]

(33). Ohtaka H; Schön A; Freire E Multidrug Resistance to HIV-1 Protease Inhibition Requires Cooperative Coupling between Distal Mutations. Biochemistry 2003, 42, 13659–13666. [PubMed: 14622012]

(34). Barrila J; Bacha U; Freire E Long-Range Cooperative Interactions Modulate Dimerization in SARS 3CLpro. Biochemistry 2006, 45, 14908–14916. [PubMed: 17154528]

(35). Sheik Amamuddy O; Verkhivker GM; Tastan Bishop O Impact of Early Pandemic Stage Mutations on Molecular Dynamics of SARS-CoV-2 Mpro. Journal of Chemical Information and Modeling 2020, 60, 5080–5102. [PubMed: 32853525]

(36). Mótyán JA; Mahdi M; Hoffka G; Tőzsér J Potential Resistance of SARS-CoV-2 Main Protease (Mpro) against Protease Inhibitors: Lessons Learned from HIV-1 Protease. International Journal of Molecular Sciences 2022, 23, 3507. [PubMed: 35408866]

(37). Khare S; Gurry C; Freitas L; Schultz MB; Bach G; Diallo A; Akite N; Ho J; Lee RT; Yeo W; Team GCC; Maurer-Stroh S GISAID’s Role in Pandemic Response. China CDC Weekly 2021, 3, 1049–1051. [PubMed: 34934514]

(38). Van Rossum G; Drake FL Python 3 Reference Manual; 2009.

(39). Katoh K; Standley DM MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. Molecular Biology and Evolution 2013, 30, 772–780. [PubMed: 23329690]

(40). Price MN; Dehal PS; Arkin AP FastTree: Computing Large Minimum Evolution Trees with Profiles Instead of a Distance Matrix. Molecular Biology and Evolution 2009, 26, 1641–1650. [PubMed: 19377059]

(41). Price MN; Dehal PS; Arkin AP FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments. PLOS ONE 2010, 5, e9490. [PubMed: 20224823]

(42). Dagum L; Menon R OpenMP: An Industry Standard API for Shared-Memory Programming. IEEE Computational Science and Engineering 1998, 5, 46–55.

(43). Team RCR: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, 2018.

(44). Xu S; Dai Z; Guo P; Fu X; Liu S; Zhou L; Tang W; Feng T; Chen M; Zhan L; Wu T; Hu E; Jiang Y; Bo X; Yu G ggtreeExtra: Compact Visualization of Richly Annotated Phylogenetic Data. Molecular Biology and Evolution 2021, 38, 4039–4042. [PubMed: 34097064]

(45). Paradis E; Schliep K Ape 5.0: An Environment for Modern Phylogenetics and Evolutionary Analyses in R. Bioinformatics 2019, 35, 526–528. [PubMed: 30016406]

(46). Wickham H Ggplot2: Elegant Graphics for Data Analysis; Springer-Verlag New York, 2016.

(47). Wang L-G; Lam TT-Y; Xu S; Dai Z; Zhou L; Feng T; Guo P; Dunn CW; Jones BR; Bradley T; Zhu H; Guan Y; Jiang Y; Yu G Treeio: An R Package for Phylogenetic Tree Input and Output with Richly Annotated and Associated Data. Molecular Biology and Evolution 2020, 37, 599–603. [PubMed: 31633786]

(48). Wickham H et al. Welcome to the Tidyverse. Journal of Open Source Software 2019, 4, 1686.

(49). Yu G Aplot: Decorate a ‘ggplot’ with Associated Information. 2022.

(50). Dowle M et al. Data.Table: Extension of ‘Data.Frame’. 2021.

(51). Wickham H; Henry L; Pedersen TL; Luciani TJ; Decorde M; Lise V Svglite: An ‘SVG’ Graphics Device. 2022.

(52). Webb B; Sali A Comparative Protein Structure Modeling Using MODELLER. Current Protocols in Bioinformatics 2016, 54, 5.6.1–5.6.37.

(53). Olsson MHM; Søndergaard CR; Rostkowski M; Jensen JH PROPKA3: Consistent Treatment of Internal and Surface Residues in Empirical pKa Predictions. Journal of Chemical Theory and Computation 2011, 7, 525–537. [PubMed: 26596171]

(54). Phillips JC; Braun R; Wang W; Gumbart J; Tajkhorshid E; Villa E; Chipot C; Skeel RD; Kalé L; Schulten K Scalable Molecular Dynamics with NAMD. Journal of Computational Chemistry 2005, 26, 1781–1802. [PubMed: 16222654]

(55). Huang J; Rauscher S; Nawrocki G; Ran T; Feig M; de Groot BL; Grubmüller H; MacKerell AD CHARMM36m: An Improved Force Field for Folded and Intrinsically Disordered Proteins. Nature Methods 2017, 14, 71–73. [PubMed: 27819658]

(56). Feller SE; Zhang Y; Pastor RW; Brooks BR Constant Pressure Molecular Dynamics Simulation: The Langevin Piston Method. The Journal of Chemical Physics 1995, 103, 4613–4621.

(57). Humphrey W; Dalke A; Schulten K VMD: Visual Molecular Dynamics. Journal of Molecular Graphics 1996, 14, 33–38. [PubMed: 8744570]

(58). Grant BJ; Rodrigues APC; ElSawy KM; McCammon JA; Caves LSD Bio3d: An R Package for the Comparative Analysis of Protein Structures. Bioinformatics 2006, 22, 2695–2696. [PubMed: 16940322]

(59). Handcock MS; Hunter DR; Butts CT; Goodreau SM; Morris M Statnet: Software Tools for the Representation, Visualization, Analysis and Simulation of Network Data. Journal of Statistical Software 2008, 24, 1–11. [PubMed: 18612375]

(60). Butts CT network: a Package for Managing Relational Data in R. J. Stat. Softw. 2008, 24.

(61). Shen Y; Gao F; Wang M; Li A RPdb: A Database of Experimentally Verified Cellular Reprogramming Records. Bioinformatics 2015, 31, 3237–3239. [PubMed: 26026167]

(62). Benson NC; Daggett V A Chemical Group Graph Representation for Efficient High-Throughput Analysis of Atomistic Protein Simulations. Journal of Bioinformatics and Computational Biology 2012, 10, 1250008. [PubMed: 22809421]

(63). Duong VT; Unhelkar MH; Kelly JE; Kim SH; Butts CT; Martin RW Protein Structure Networks Provide Insight into Active Site Flexibility in Esterase/Lipases from the Carnivorous Plant Drosera Capensis. Integrative Biology 2018, 10, 768–779. [PubMed: 30516771]

(64). Seidman SB Network Structure and Minimum Degree. Social Networks 1983, 5, 269–287.

(65). Butts CT Social Network Analysis with sna. J. Stat. Softw. 2008, 24.

(66). Smith EC; Denison MR Implications of Altered Replication Fidelity on the Evolution and Pathogenesis of Coronaviruses. Current Opinion in Virology 2012, 2, 519–524. [PubMed: 22857992]

(67). Domingo E; Perales C Viral Quasispecies. PLOS Genetics 2019, 15, e1008271. [PubMed: 31622336]

(68). Metcalf CJE; Birger RB; Funk S; Kouyos RD; Lloyd-Smith JO; Jansen VAA Five Challenges in Evolution and Infectious Diseases. Epidemics 2015, 10, 40–44. [PubMed: 25843381]

(69). Frost SDW; Pybus OG; Gog JR; Viboud C; Bonhoeffer S; Bedford T Eight Challenges in Phylodynamic Inference. Epidemics 2015, 10, 88–92. [PubMed: 25843391]

(70). Morel B; Barbera P; Czech L; Bettisworth B; Hübner L; Lutteropp S; Serdari D; Kostaki E-G; Mamais I; Kozlov AM; Pavlidis P; Paraskevis D; Stamatakis A Phylogenetic Analysis of SARS-CoV-2 Data Is Difficult. Molecular Biology and Evolution 2021, 38, 1777–1791. [PubMed: 33316067]

(71). Moustaqil M; Ollivier E; Chiu H-P; Van Tol S; Rudolffi-Soto P; Stevens C; Bhumkar A; Hunter DJB; Freiberg AN; Jacques D; Lee B; Sierecki E; Gambin Y SARS-CoV-2 Proteases PLpro and 3CLpro Cleave IRF3 and Critical Modulators of Inflammatory Pathways (NLRP12 and TAB1): Implications for Disease Presentation across Species. Emerging Microbes & Infections 2021, 10, 178–195. [PubMed: 33372854]

(72). Ceraolo C; Giorgi FM Genomic Variance of the 2019-nCoV Coronavirus. Journal of Medical Virology 2020, 92, 522–528. [PubMed: 32027036]

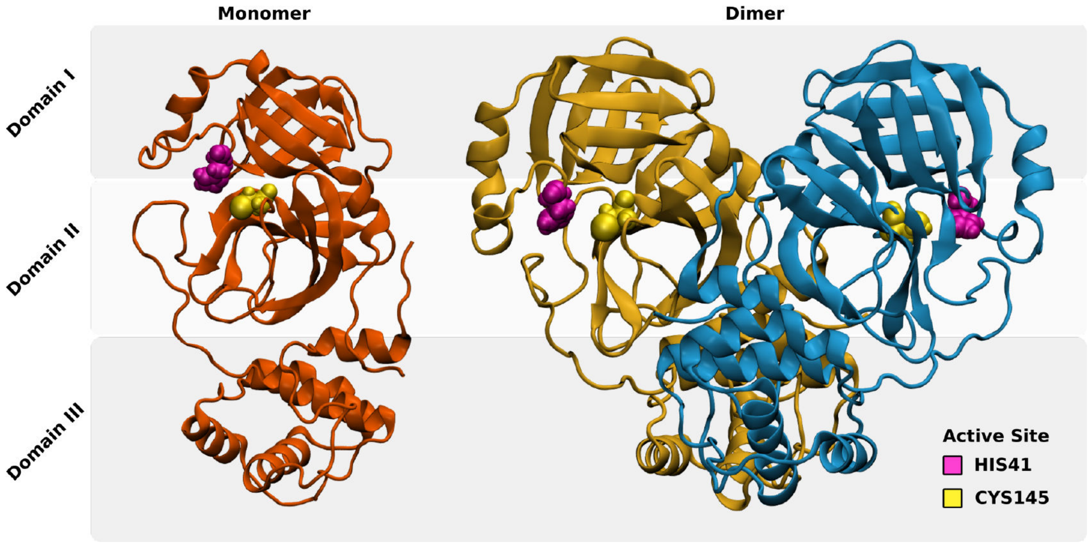


Figure 1:

Monomer and dimer conformations of the wild-type SARS-CoV-2 main protease (M^{pro}), based on respective atomistic molecular dynamics simulations of the free monomer (left) and dimer (right); MD simulations were based on the 6Y2E PDB crystal structure of M^{pro} [12], as described in the Methods section. Note the three domains (highlighted, left); the active site straddles the cleft between domains I and II, and faces away from the dimerization interface.

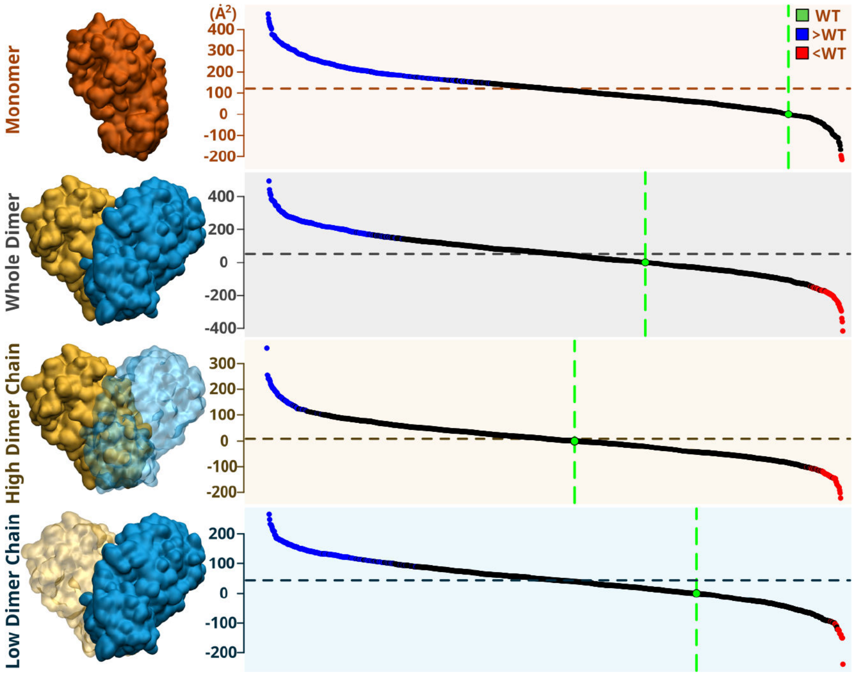


Figure 2:

Total mean SASA distribution of the monomer, dimer, and each dimer chain, across variants. WT value is in green; trajectories significantly higher than WT are shown in blue, lower in red (black values do not differ significantly from WT). Substantially more variants show increased SASA versus WT than decreased SASA. This is particularly true for free monomers, suggesting that mutations act in part through modifications to interfacial surface that is buried in the dimer.

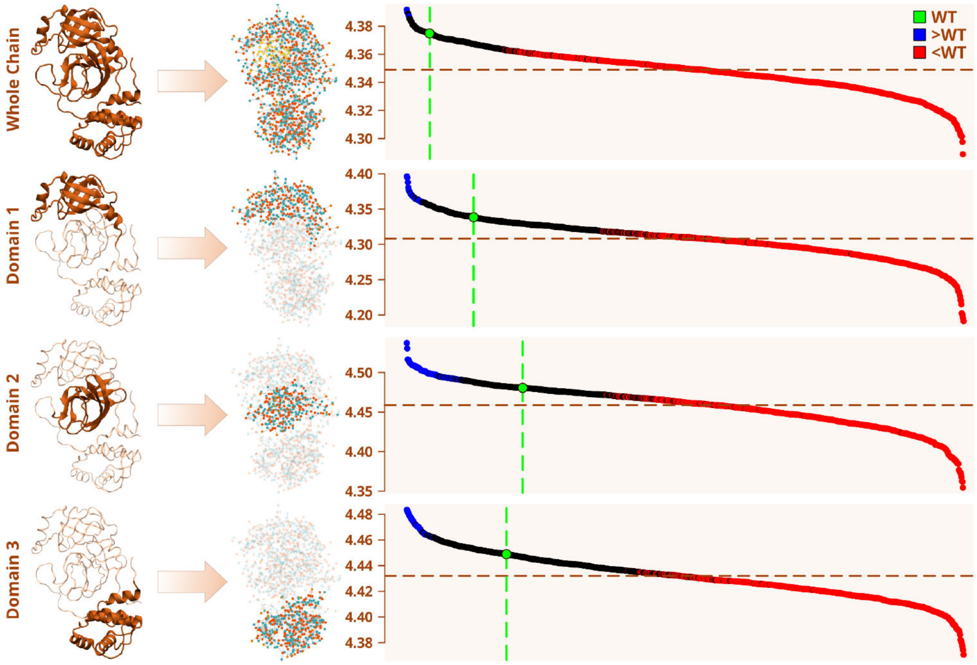


Figure 3:

Mean cohesion of variants in the monomeric conformation, in decreasing order. WT is highlighted in green. Variants with mean cohesion scores significantly greater than WT are colored blue, and those significantly less than WT are colored red. A horizontal line through the distribution marks the grand mean. The majority of variants show less cohesion both for the monomer as a whole, and in each domain.

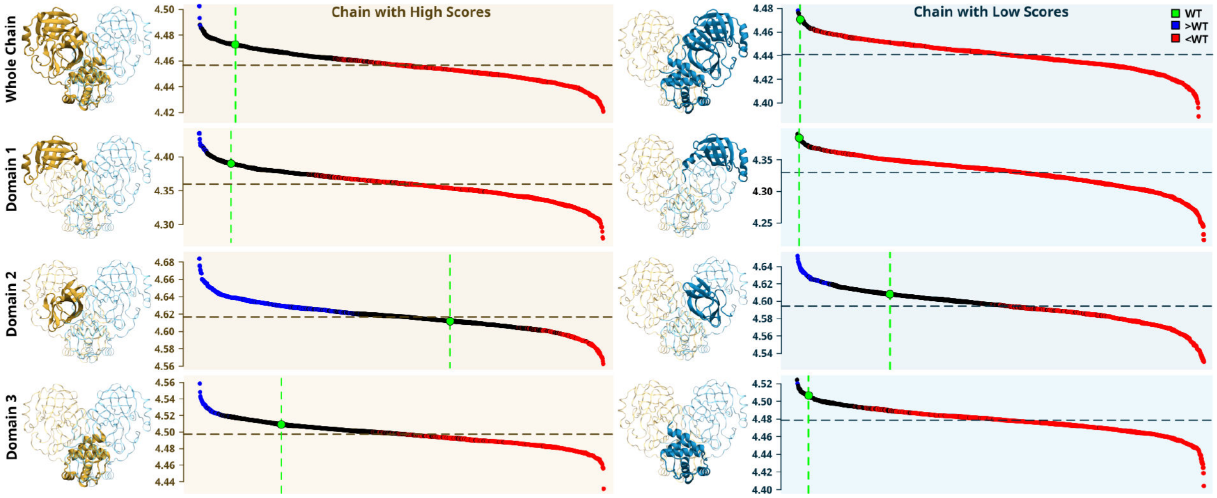


Figure 4:

Mean cohesion values for of all variants in the dimeric state. The same plot style and color scheme are used as in Figure 3. To break homodimer symmetry, chains were labeled for analysis based on the observed mean cohesion score (left higher, right lower).

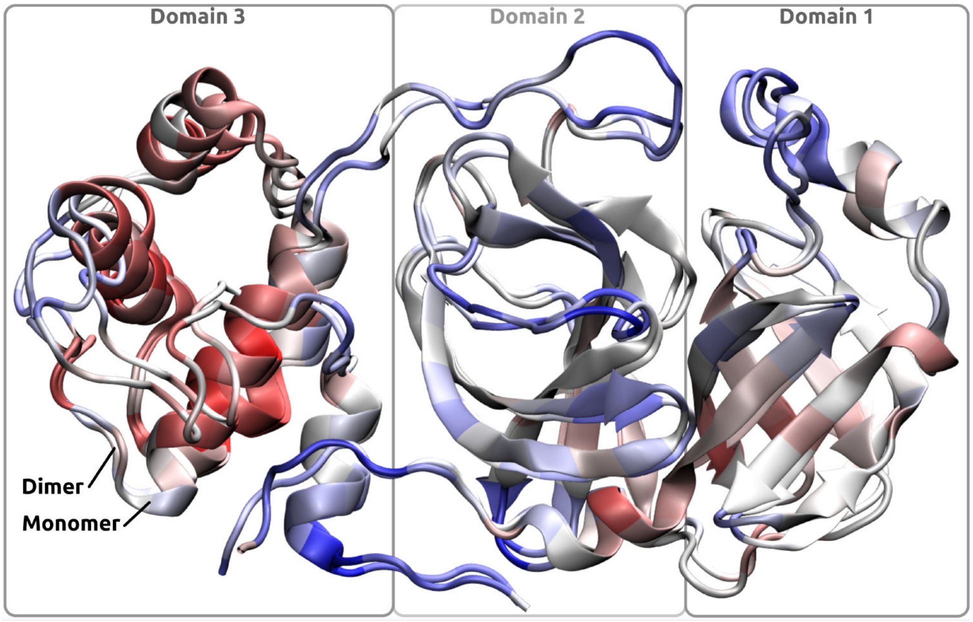


Figure 5:

Comparison of monomer and dimer structures, with coloring corresponding to the log-ratio of between-chain variance and within-chain variance. Blue color shows higher between-chain variance, red shows higher within-chain variance. Free monomer and dimeric monomer structures are overlaid; both show very similar patterns of change in backbone torsion angles.

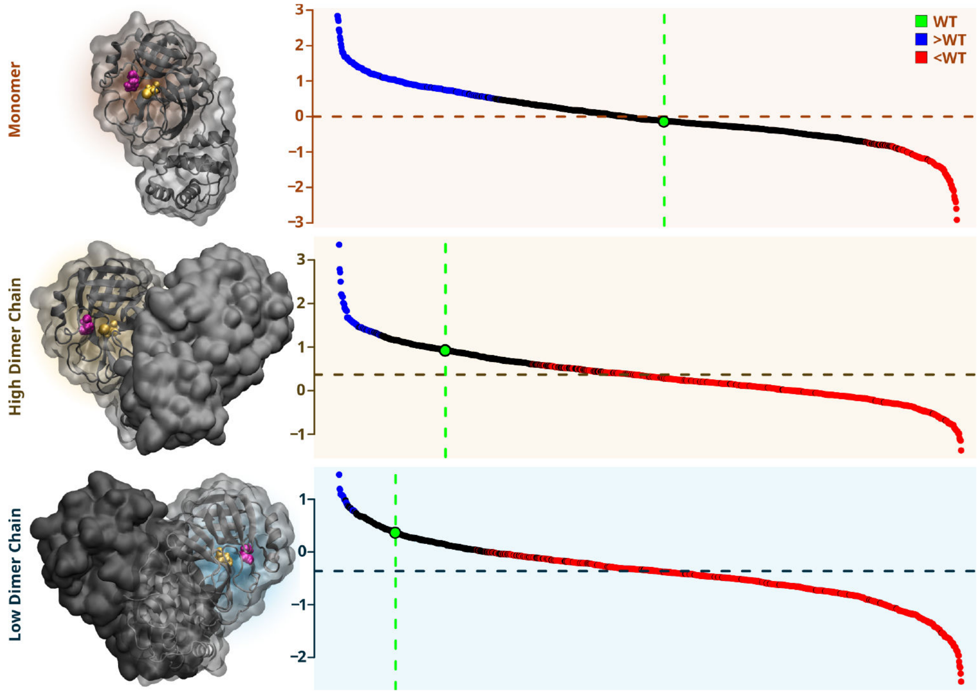


Figure 6:

Mean ASN constraint scores by variant trajectory, for free monomeric and dimeric states; to break symmetry, dimeric active sites labeled based on mean constraint for analysis (middle high, bottom low). WT values indicated in green, grand mean indicted by horizontal line. Blue values are significantly more constrained than WT, red values are significantly less, black values not significant. Dimer active sites show reduced constraint for most variants, with no trend for the free monomer.

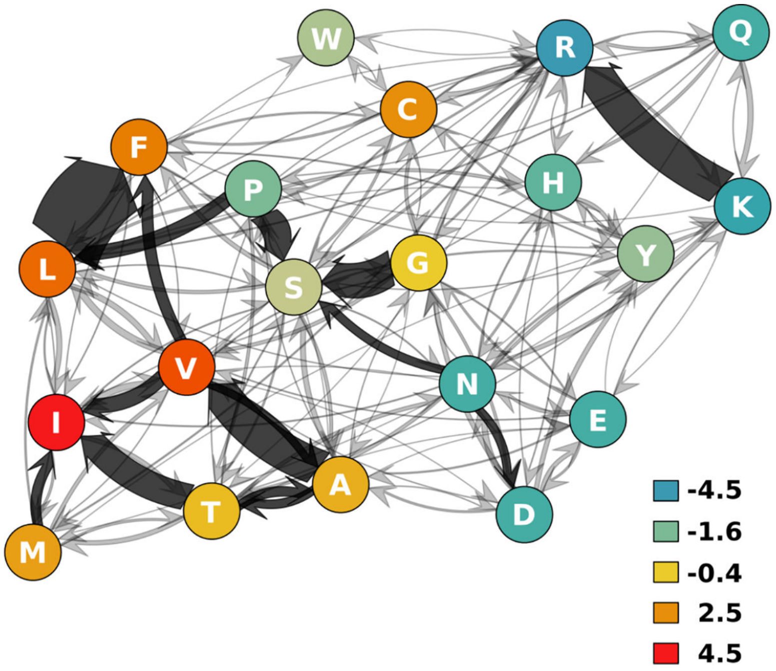


Figure 7:

Substitution network showing the trends in residue substitutions. Nodes represent unique amino acids, with directed edges in the direction of the substitution. Edges are weighted by the number of substitutions observed, with darkened edges for substitutions which occurred more than 20 times. Nodes are colored by the corresponding hydrophobicity of the amino acid.

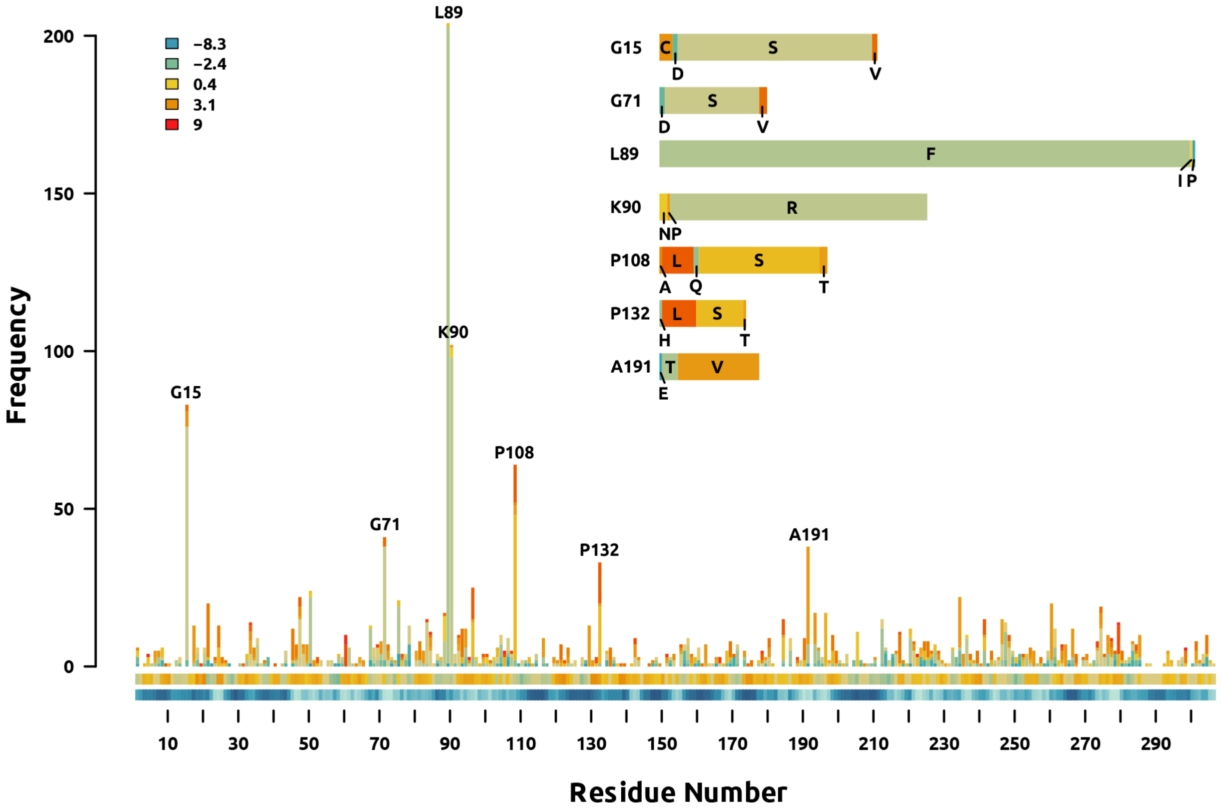


Figure 8:

Frequency of substitutions along the main protease sequence. Colors indicate change in hydrophobicity resulting from the substitution, ranging from decreased hydrophobicity (blue) to increased hydrophobicity (red). A rectangular moving average of mean hydrophobicity change is shown below the bar plot using the same color scale. A rectangular moving average of the mean RSA of residues in the dimer conformation is shown at the bottom of the plot; darker values correspond to a more buried residue.

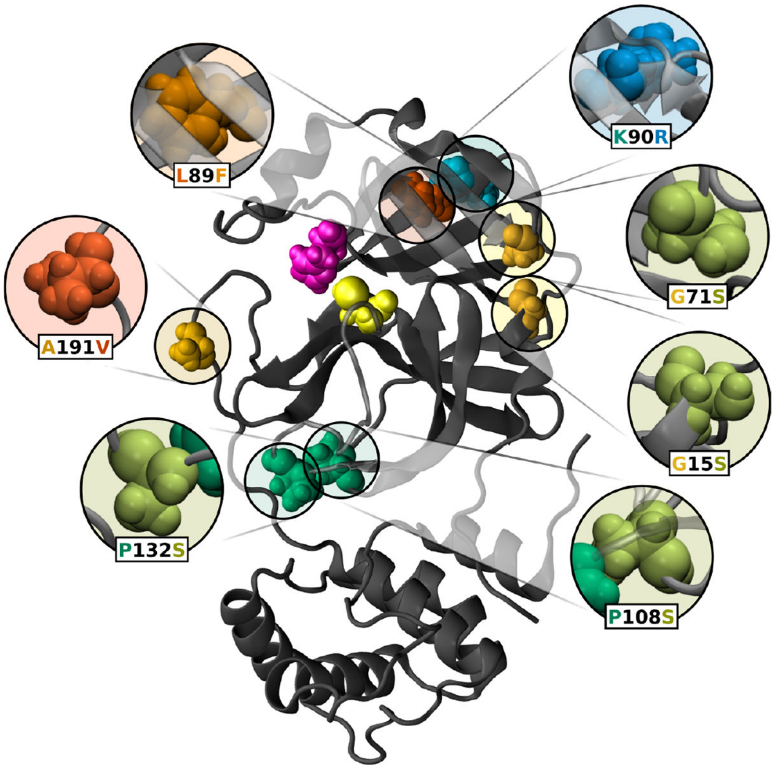
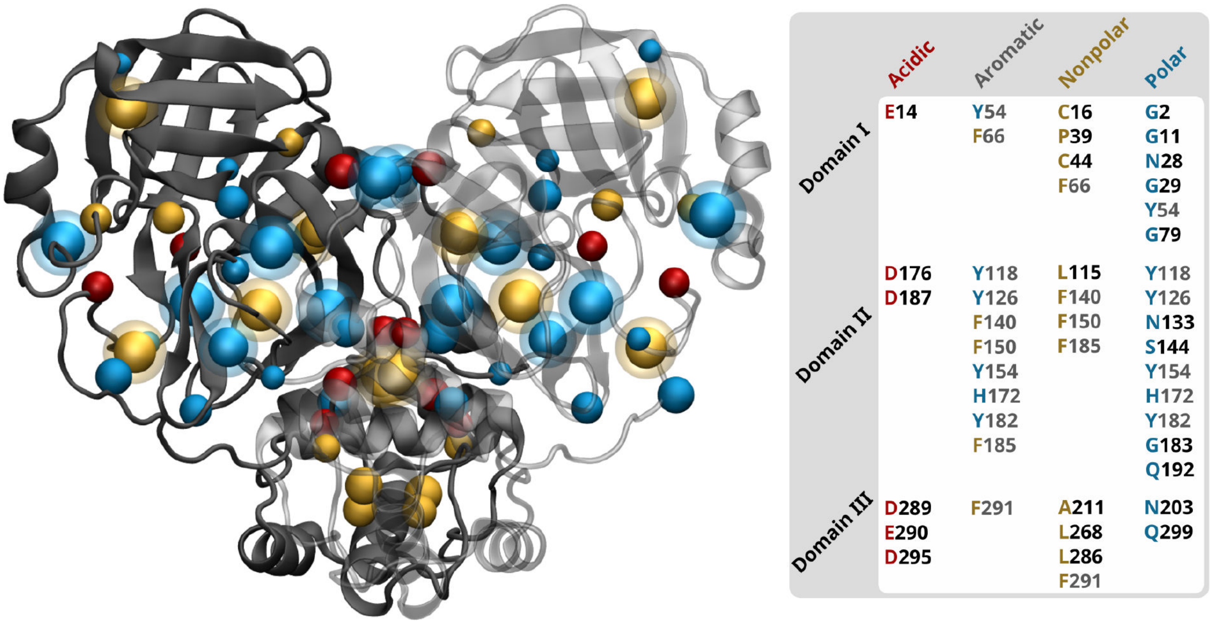


Figure 9:

Locations of persistent substitutions in a single chain of the main protease structure are shown by the respective amino acid vdW representation colored according to hydrophobicity, as well as the catalytic C145 and H41.



	Acidic	Aromatic	Nonpolar	Polar
Domain I	E14	Y54 F66	C16 P39 C44 F66	G2 G11 N28 G29 Y54 G79
Domain II	D176 D187	Y118 Y126 F140 F150 Y154 H172 Y182 F185	L115 F140 F150 F185	Y118 Y126 N133 S144 Y154 H172 Y182 G183 Q192
Domain III	D289 E290 D295	F291	A211 L268 L286 F291	N203 Q299

Figure 10:

Conserved residues visualized in VMD using beads in their location on the dimer structure. Residues with a “halo” have an aromatic side-chain (Tyr, Phe, Hse). Blue are polar (Tyr, Asn, Gln, Ser, Hse, Gly), yellow are nonpolar (Phe, Cys, Ala, Leu, Pro). Acidic residues (Asp, Glu) are colored red.

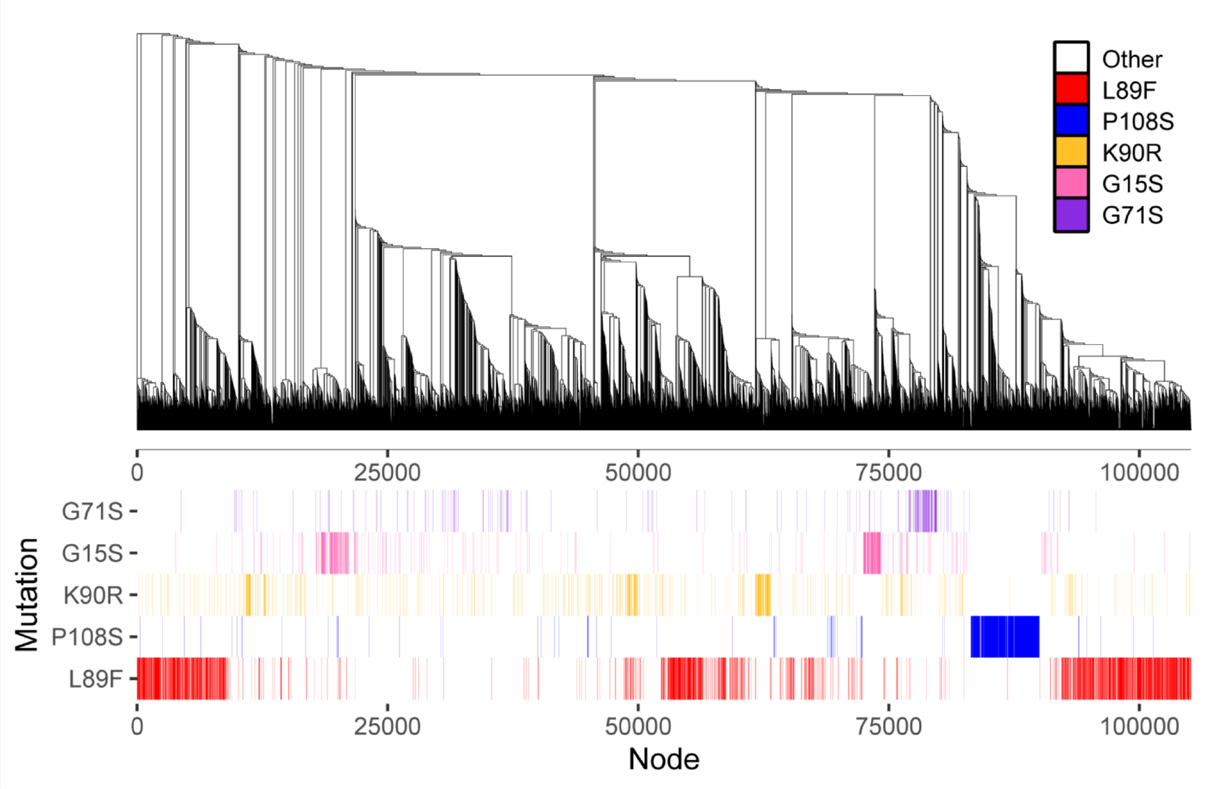


Figure 11:

Phylogenetic tree (topology only) generated using all available full genomes from 1,253 M*pro* variants as of February 25, 2021, including variants with multiple non-synonymous mutations, and one WT reference sequence[13]. The five most common mutations are indicated by colored lines: purple - G71S, pink - G15S, orange - K90R, blue - P108S, red - L89F.