

responses to existing therapies.^{15,16} Nevertheless, coronaviruses comprise an RNA-dependent RNA polymerase genetic proof-reading machinery,^{17,18} which may be responsible for their reduced genetic diversity, even though the persistence of the pandemic may allow for the accumulation of immunologically important mutations.³ From the data gathered from the GISAID database¹² and real-time subsample estimates of genetic relatedness from the Nextstrain web resource,¹⁹ it is clear that the virus is evolving within the human host. More recently, a SARS-CoV-2 spike variant was reported to occur asynchronously in multiple geographic locations with the capacity for expressing higher viral concentrations in the upper respiratory tract,³ hinting to some form of selection.

Although progress is being gradually made in understanding the viral structural biology and symptomatology of the disease, current knowledge is still fragmentary.^{20,21} The death toll and the number of infections keep on rising albeit at variable rates in different parts of the world. Thus, time is of the essence for the discovery of effective therapies. It is imperative to better characterize parts of the viral mechanisms to better understand the behavior of the new coronavirus. Already, with the help of experimentally determined structures, genomic data, and annotations, a growing number of *in silico* work suggests potential solutions to the COVID-19 pandemic using various techniques, including the use of molecular docking^{22–27} and dynamics (MD),^{28–36} network analysis,^{37–43} and machine learning.^{2,44,45} Collectively, these may pave the way to a potential solution by prioritizing inhibitors or drug targets against COVID-19. Most of the mentioned MD research has been focused on the application of the technique to filter and prioritize hit compounds from small compound docking. Suárez and Díaz, however, have investigated apo and substrate-bound M^{Pro} using 2 μ s MD simulations in monomeric and dimeric forms, to reveal instability of domain III in the former, using 4 protease crystal structures from SARS-CoV-2.³⁶ In this study, we examine some of the early mutations of SARS-CoV-2 M^{Pro} and investigate different aspects of their dynamics using MD.

The M^{Pro} enzyme, also known as the 3C-like protease, is one of the best studied drug targets among the coronaviruses.⁴⁶ This is mainly due to the similarities in their active site and their mechanisms with the related pathogenic betacoronaviruses from previous epidemics of SARS-CoV and MERS-CoV (Middle East respiratory syndrome coronavirus).⁴⁷ M^{Pro} is a conserved drug target present in all members of the *Coronavirinae* subfamily^{48,49} and is highly similar to its SARS-CoV-2 counterpart.⁴⁷ SARS-CoV-2 M^{Pro} does not have a human homolog,³⁹ which reduces the chances of accidentally targeting host proteins. Alongside the papain-like protease (PLP) enzyme, M^{Pro} plays an essential role in the process of viral maturation,² cleaving the large precursor replicase polyprotein 1ab to produce 16 non-structural proteins.^{2,50} The cysteine protease functions as a homodimer and mainly comprises three domains (I–III).² Homodimerization plays an important role in the catalytic activity of M^{Pro}, as reported in the case of the SARS-CoV M^{Pro} homolog, where the G11A mutation completely abolished its activity by interfering with the insertion of the N-finger region (residues 1–9).⁵¹ Additionally, only one of the dimers was shown to be active at a time in⁵² SARS-CoV M^{Pro}. At the N-terminus, the chymotrypsin-like domain I (residues 10–99) is connected to the picornavirus 3C-protease-like domain II (residues 100–182), which together form a hydrophobic substrate binding

site, with catalytic residues HIS41 and CYS145.^{50,53} Domain III (residues 198–303), also referred to as the helical domain, is connected to domain II⁵⁴ by a 15-residue linker loop and was shown to regulate enzymatic activity in SARS-CoV.⁵⁵ While each domain minimally makes contact with its equivalent domain from the alternate chain, the majority of the dimer contact interface is a result of interactions present between domain II (chain A) and the N-finger (chain B).⁵⁰ In the same manner, the N-finger from chain A makes contact with domain II from chain B. Each chain is referred to as a protomer.^{34,42,50}

In this work, we study the collective effects of various M^{Pro} mutations from a filtered sample of 50 SARS-CoV-2 isolates by first mapping them on 3D structures and performing all-atom MD for each of the mutants, in addition to the reference protein. All-atom simulations were carried out at a constant protonation state initially corresponding to a pH of 7. Multiple aspects of the protein dynamics were analyzed using a battery of techniques, including the averaged betweenness centrality (BC)—a metric of dynamic residue network (DRN) analysis, dynamic cross-correlation (DCC),⁵⁶ geometry calculations (interdomain angles and interprotomer distances) based on the center of mass (COM), cavity compaction analyses, the anisotropic network model (ANM), and the analysis of residue and backbone fluctuations. A new metric based on Jaccard's coefficient was developed to sort non-Gaussian distributions to facilitate the differentiation of large numbers of such curves obtained from MD reaction coordinates. Coarse-grained Monte Carlo simulations were also independently investigated.

2. MATERIALS AND METHODS

2.1. Sequence and Template Retrieval. A high-resolution (1.48 Å) biological unit for crystal structure of the SARS-CoV-2 main protease (PDB ID 5RFV⁵⁷) was retrieved from the Protein Data Bank (PDB)⁵⁸ to be used as a template for homology modeling. Its sequence was used as a reference throughout this work. PyMOL (version 2.4)⁵⁹ was used to remove any non-protein molecules and to reconstitute the biological unit as chains A and B. SARS-CoV-2 genomes of any length were acquired from the GISAID website as a FASTA-formatted file.¹²

2.2. Building the Mutation Data Set. Low coverage sequences (entries with >5% unknown nucleotides) from GISAID were not selected. A local BLAST database was then set up for these sequences using the *makeblastdb* command available from the BLAST+ application (version 2.8.1).⁶⁰ Protease mutants were subsequently retrieved using the reference sequence as a query parameter for the *tblastn* command with default parameters, except for the maximum number of target sequences, which was set at 10000. Identical sequences were then filtered out before selecting BLAST hits that had a 100% sequence coverage and a percentage sequence identity of <100%. Sequences coming from non-human hosts were discarded. In order to retain as much data as possible and minimize the incorporation of sequencing errors, fold coverage was used where provided, while quality was imputed on the basis of an identical M^{Pro} sequence being present more than once in the filtered data set for samples not possessing sequencing coverage information. This resulted in 50 mutant sequences with either high coverage, where available, or with additional sources of support otherwise. Further details about the samples are given in the [Supporting Information acknowledgment, Table S1](#).

2.3. Homology Modeling, pH Adjustment, and Analysis of Residue Interactions. PIR-formatted target-template sequence alignment files were generated for each mutant using the BioPython library (version 1.76)⁶¹ within ad hoc Python scripts for use in MODELLER (version 9.22).⁶² The *automodel* class was used with slow refinement and a deviation of 2 Å to generate 12 models in parallel for each mutant, after which the ones with the lowest z-DOPE scores were retained. The protein was then adjusted to a pH of 7 using the PROPKA algorithm from the PDB2PQR tool (version 2.1.1).⁶³ For visualizing the overall interactions at given residue positions, the Arpeggio tool⁶⁴ was used to programmatically generate the inter-residue interactions before computing their sums using an in-house Python script. More generally, the Discovery Studio Visualizer (version 19.1) was used for describing the non-bonded interactions.⁶⁵

2.4. Molecular Dynamics Simulations. All-atom protein MD simulations (using the AMBER03 force field) were run for the protonated dimers using GROMACS (version 2016.1)⁶⁶ at the Center for High Performance Computing (CHPC). Proteins were placed in a triclinic box containing 0.15 M NaCl in SPC-modeled water. A minimum image distance of 1.5 nm between the solute and the box was used. The system was then energy minimized using the steepest descent algorithm with an initial step size of 0.01 nm for a maximum force of 1000 kJ/mol/nm and a maximum of 50,000 steps. Temperature was subsequently equilibrated at 310 K for 50 ps, according to the *NVT* ensemble. Pressure was then equilibrated at 1 bar for 50 ps, using the Berendsen algorithm according to the *NPT* ensemble. During both *NVT* and *NPT*, the protein was position restrained, and constraints were applied on all bonds. Unrestrained production runs (100 ns) were then performed, where constraints were applied only on H-bonds, and the Parrinello–Rahman algorithm was used for pressure coupling. In all cases, a time step of 2 fs was used, with a short-range non-bonded cutoff distance of 1.1 nm and the PME algorithm for long-range electrostatic interaction calculations. Prior to analysis, the periodic boundary conditions were removed, and the trajectories were corrected for rotational and translational motions.

2.5. Coarse-Grained Simulations. Coarse-grained (CG) models enable simulations of long timescales for protein systems and assemblies and represent a computationally effective strategy for adequate sampling of the conformational space while maintaining physical rigor. The CABS model was employed for multiple CG simulations^{67–71} of the SARS-CoV-2 main protease dimer structures (PDB IDs SRFV, 6Y2E, and 6Y2F⁵⁰). In this model, the CG representation of protein residues is reduced to four united atoms. The residues are represented by main-chain α -carbons (C_α), β -carbons (C_β), the COM of side chains, and another pseudoatom placed in the center of the C_α – C_α pseudo-bond. The sampling scheme involved Monte Carlo (MC) dynamics moves including local moves of individual residues and moves of small fragments composed of 3 protein residues. One hundred independent CG simulations were carried out for each studied system with the CABS-flex standalone Python package for fast simulations of protein dynamics, which is implemented as a Python 2.7 object-oriented package.⁷¹ In each simulation, the total number of cycles was set to 1000, and the number of cycles between trajectory frames was 100. Accordingly, the total number of generated models was 2,000,000, and the total number of saved models in the trajectory used for analysis was

20,000. It was previously shown that the CABS-flex approach can accurately recapitulate all-atom MD simulations on a long timescale.^{67–71} The results of 100 independent CG-CABS simulations for each system were averaged to obtain adequate sampling and ensure convergence of simulation runs.

2.6. Dynamic Residue Network Analysis, Dynamic Cross-Correlation, and Normal Mode Analysis. The MD-TASK tool kit⁵⁶ was used to calculate the residue averaged *BC* values over the last 50 ns of simulation for each protein using a cutoff distance of 6.70 Å and a step size of 25, generating a total of 10,001 frames. DCC was calculated for each of the proteins using the same frames and time steps before linearizing each matrix. Pairwise Pearson correlations were then performed for all linearized matrices before performing hierarchical clustering. In all cases, the C_β and glycine C_α atoms were used. The GROMACS commands *trjconv* and *make_ndx* were utilized to reduce the trajectory sizes, to only keep C_α and C_β atoms prior to computation. Normal modes were obtained from C_α using ProDy,⁷² with default parameters for the ANM.

2.7. Pocket Detection and Dynamic Analysis. The reconstituted biological unit for the reference structure was submitted to the FTMap web server⁷³ using default parameters. The PyMOL plugin PyVOL⁷⁴ was then used to identify the surfaces of any potential cavity, specifying the protein as selection, with a default minimum volume of 200 Å³. Predictions from both tools were combined. Residues for the interprotomer subpocket were defined by visually inspecting residues in proximity to the cavity surfaces detected by PyVOL that overlapped with part of the FTMap probe binding predictions. The pocket was selected due to its location and accessibility to the outside. The substrate binding residues (identified using PyVOL) from both protomers of M^{Pro} were also investigated due to their functional importance in catalysis, even though FTMap did not identify a binding hot spot at that location. The radius of gyration (R_g) for the subpocket and the binding pockets was then computed from the entirety of the simulated MD data in each case, using the GROMACS *gyrate* command. The generated data was then visualized and analyzed using various open source Python libraries, such as matplotlib,⁷⁵ Seaborn, Pandas,⁷⁶ NumPy,⁷⁷ SciPy,⁷⁸ MDTraj,⁷⁹ and NGLview.⁸⁰

2.8. Coupling Peak Finding to an Adapted Jaccard Distance Metric to Rank and Extract Equilibrium Conformations from KDE Distributions of MD Measurements. In order to facilitate the comparison of distributions of kernel density estimation (KDE) from a large number of simulations, a new metric based the Jaccard distance (d_j) was devised and applied to rank both Gaussian and non-Gaussian distributions obtained from MD measurements based on their distance from a comparator. Non-Gaussian curves can generally be observed from a collective variable when multiple equilibrium states are sampled (leading to multimodal distributions) or when there is a skewed distribution. The metric assumes no similarity of variance or distribution shapes but requires a range [x_{\min} , x_{\max}] common to both samples to be set for the KDE fitting calculations and a common area under the curve, here having a value of one. KDE d_j has a minimum of zero and a maximum of one, corresponding to complete dissimilarity and identity, respectively. Scott's rule, defined by the equation ($n^{-1/d+4}$), was used for Gaussian kernel bandwidth selection, where n and d are the sample size and

the number of dimensions.^{78,81} The fitting range was set to the global minimum and maximum values of the complete data set to be compared. The equation used, which is based on numerical approximation of the integrals, is defined as

$$\text{KDE } d_j = 1 - \frac{\sum_{x_{\min}}^{x_{\max}} \sum_{g(x) \geq f(x)} f(x) + \sum_{x_{\min}}^{x_{\max}} \sum_{f(x) > g(x)} g(x)}{\sum_{x_{\min}}^{x_{\max}} \sum_{f(x) \geq g(x)} f(x) + \sum_{x_{\min}}^{x_{\max}} \sum_{g(x) > f(x)} g(x)} \quad (1)$$

where $f(x)$ and $g(x)$ are the kernel density estimates of the reference and test samples. The metric was applied to rank the mutant distributions according to their distance from the reference M^{pro} sample in the case of C_{α} root-mean-square deviation (RMSD) and COM distance distributions.

Thereafter, a straightforward 1D peak finding algorithm was applied to detect the modes from the ranked samples. The KDE modes correspond to zones in a conformational landscape about which structures tend to visit more frequently. Only peaks above a set minimum KDE estimate ($\epsilon = 0.05$) were used to limit the detection of insignificantly low peaks. Using this approach, equilibrium conformations corresponding to various energy minima along given reaction coordinates were extracted.

3. RESULTS AND DISCUSSION

3.1. Analysis of Residue Mutations and Their Distribution in the 3D Structure. As a preliminary investigation of the propensity of the sequences to acquire novel mutations, unique residue mutations were determined across our set of 50 protein sequences, which were filtered from the GISAID database.¹² While we cannot infer population frequencies (across the world) from our relatively small sample, we show from our estimate that multiple missense mutations have already occurred on each domain of the M^{pro} (Figure 1). These mutations are shown in Table 1, and further experimental support is reported in the Supporting Information, Table S1. From these, it can be observed that many mutations are interconversions of the hydrophobic side-chain residues alanine and valine. As seen in Figure 1, most residue mutations have occurred in solvent-accessible surfaces, with the exception of A7V, V20L, L89F, A116V, A129V, T135I, I136V, V157I/L, C160S, A173V, T201A, A234V, and A266V, which were predicted to be buried by the PyMOL script *findSurfaceResidues*, using a default cutoff of 2.5 Å².

Current genomic sequence data suggests a generally low rate of mutation in SARS-CoV-2.^{82,83} This being said, in the case of the M^{pro} enzyme, a relatively higher rate of non-synonymous mutations has occurred at residue position 15 (G15D/S) in domain I, residue position 157 (V157I/L) in domain II, and at position 184 (P184L/S) within the interdomain linker region. On the 3D structure, it can be seen that these mutable areas occur away from the core areas of domains I and II, hence the probable lower selective pressure for these regions. For this reason, we posit that individually, these loci may be less important for basic enzymatic functions. Additionally, samples EPI_ISL_425242 and EPI_ISL_419710 have accumulated two mutations each. Of notable interest is the mutation of the active site flap residue 48 from the “TSEDMLN” loop described by Ma and coworkers⁸⁴ to “TSEEMLN” in sample EPI_ISL_425242. Earlier, the homologous region displayed the motif “TAEDMLN” in SARS-CoV. It is possible that this may affect the affinity of the enzyme for its substrate and play a

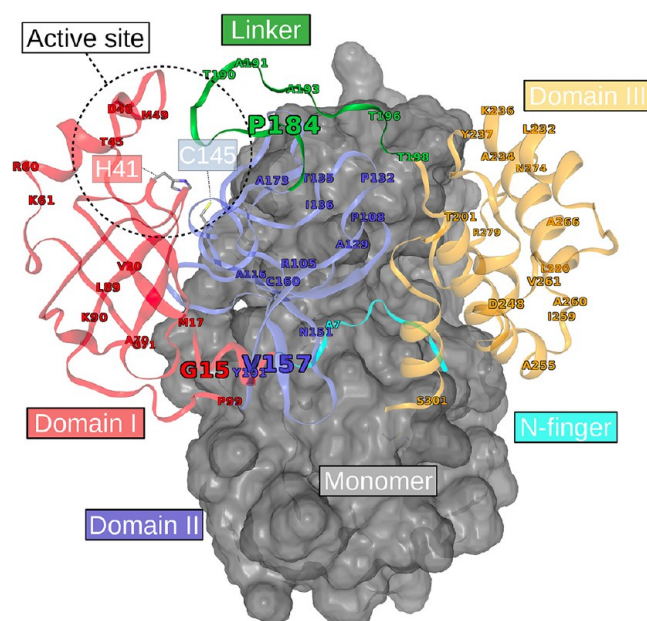


Figure 1. Mapping of the positions showing unique mutations from the reference M^{pro} sequence. For clarity, domains (I–III) are colored (red, blue, and orange, respectively) only for one of the monomers, while the other is represented as a gray surface. The domain linker region is in green, and the N-finger is in cyan. The size of the labels denotes the number of unique mutations recorded at that position.

role in the virus’ level of fitness. Residue 15 mutations are examined in section 3.2.

Mutation rates of RNA viruses are known to be generally high, endowing them with the ability to escape host immune responses, improve their virulence, and even change tissue tropism.^{14,85} Extinction events are not uncommon among the RNA viruses, as seen in the influenza A H1N1 strains⁸⁶ and the previous SARS-CoV strain,⁸⁷ and may be associated with the gradual accumulation of non-synonymous mutations. On the other hand, a reduction in replicative speed and genetic diversity was independently observed in the poliovirus 3D^{G64S} mutant compared to its wild-type (WT) when rates of mutation were artificially increased by exposure to a mutagen.^{14,88,89} In the case of the HIV, multiple mutations of minor effect are known to collaboratively modulate the effect of other advantageous mutations already present in the protease.^{90,91} As a newly emerged pathogen with a relatively long incubation period and an incompletely understood biology, these facts from related viruses presuppose a potentially complex mechanism of viral evolution and adaptation, which suggests that mutations have to be closely monitored for global health and security. It is interesting to note that among the buried residue mutations, domain II has accumulated the highest number of these mutations in such little time, which may suggest a certain degree of tolerance to mutations in that region, despite their presence within beta strands. In the same domain, the A116V mutation occurs on a beta strand, which is supported by a rich network of hydrogen bonds. The local impacts of the A116V mutation are discussed further in section 3.2.

While several mutations are present in domain III, the current data suggests that these have not yet (at the time of writing) further evolved at these positions and probably suggests that there might have a higher fitness cost involved

Table 1. List of Missense Mutations in SARS-CoV-2 M^{Pro}^a

GISAID ID	mutation	GISAID ID	mutation	GISAID ID	mutation
EPI_ISL_425319*	A7V	EPI_ISL_425284*	A116V	EPI_ISL_424470*	T196M
EPI_ISL_420422	G15D	EPI_ISL_421005*	P108S	EPI_ISL_423642	T201A
EPI_ISL_420181	G15S	EPI_ISL_422860*	A129V	EPI_ISL_419256*	L220F
EPI_ISL_425242	G15S, D48E	EPI_ISL_420579	P132L	EPI_ISL_421506	L232F
EPI_ISL_423772*	M17I	EPI_ISL_425655	T135I	EPI_ISL_425235	A234V
EPI_ISL_425342	V20L	EPI_ISL_420182	I136V	EPI_ISL_426097*	K236R
EPI_ISL_421312*	T45I	EPI_ISL_420510*	N151D	EPI_ISL_416720*	Y237H
EPI_ISL_425839*	M49I	EPI_ISL_415503	V157I	EPI_ISL_425886*	D248E
EPI_ISL_418269*	R60C	EPI_ISL_426028	V157L	EPI_ISL_418075	A255V
EPI_ISL_420306	K61R	EPI_ISL_417413	C160S	EPI_ISL_422919	I259T
EPI_ISL_421763*	A70T	EPI_ISL_418082	A173V	EPI_ISL_423725*	A260V
EPI_ISL_413021	G71S	EPI_ISL_423288	P184S	EPI_ISL_425498*	V261A
EPI_ISL_415643	L89F	EPI_ISL_420241*	P184L	EPI_ISL_421380*	A266V
EPI_ISL_420059	K90R	EPI_ISL_419710*	A191V, L220F	EPI_ISL_420610*	N274D
EPI_ISL_419756	P99L	EPI_ISL_415610*	A193V	EPI_ISL_425643	R279C
EPI_ISL_425132	Y101C	EPI_ISL_421515*	T198I	EPI_ISL_422184*	S301L
EPI_ISL_419984*	R105H	EPI_ISL_423007	T190I		

^aSamples with noticeably large differences in C_α RMSD from the reference protease are marked with an asterisk.

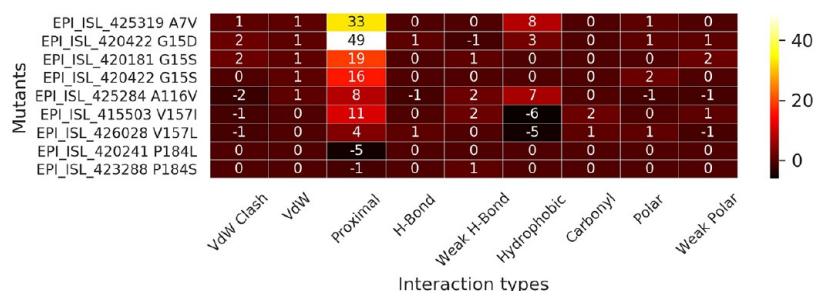


Figure 2. Differences in the sum of each interaction type for the M^{Pro}, only for the residue locations that either accumulated more than one non-synonymous mutation or the ones occurring close to protein interfaces. The differences were obtained by subtracting the reference values from the matching residue loci in the mutant. Sample names and the selected residue mutations are shown along the y-axis.

with mutating residues from this domain. This may be due to its major role in regulating enzyme activity, together with the stabilizing N-finger region.⁵³ It is also possible that under-sampling may have led to a similar observation. Most of the mutations have occurred in solvent-exposed surfaces of the protein, which may be under reduced selective pressure, especially if these are in loop regions. On the other hand, interfacial residue mutations (particularly at the chain interface) may exert their effects in a more exacerbated manner by either overstabilizing or destabilizing protein–protein interactions, as reported in works on other proteins.^{92,93} One such mutation has already happened at position 7 (A7V) in the N-finger. The local impacts of the A7V mutation are discussed in section 3.2.

3.2. Homology Modeling and Inspection of Residue Protonation States. After the preliminary analysis of M^{Pro} at the sequence level, their 3D structures were built for further investigation. The z-DOPE scores for the best homology models obtained for each sample were all below -1 , indicating that they were all native-like.^{62,94} Overall, z-DOPE values had a minimum of -1.48 , a maximum of -1.38 , with a median of -1.42 . While it is not possible to simulate changes in the protonation state using classical MD (especially when there are many titratable residues), an initial approximation of the most prevalent residue protonation states (at pH 7) was used for each of the M^{Pro} samples. As there was a relatively higher level of variation in residue protonation states among each of the

seven histidine residues found in each M^{Pro} protomer, only the catalytic residue and the missense mutations are described herein, post homology modeling. We suspect that the high number of titratable amino acids may play a role in influencing protein behavior at varying pH levels. Previous work in SARS-CoV indeed reported a pH-dependent activity switch of the main protease.⁹⁵ In our case, the catalytic residue HIS41 was generally protonated at the delta nitrogen (HID) atom but also occurred in its fully protonated state (HIP) in one of the protomers for samples EPI_ISL_419710 and EPI_ISL_425655. Protonated aspartic acid (ASH) was found in both protomers of sample EPI_ISL_420510, which was the only isolate to contain the N151D mutation. ASH was also found in sample EPI_ISL_421312, for only one of its protomers at residue position 289, which otherwise occurs in its deprotonated form in all other samples. The R105H mutation (present only in sample EPI_ISL_419984) occurs as an HID in each protomer. Similarly, the Y237H mutation, present only in sample EPI_ISL_416720, occurs as an HID in both of its protomers. The local residue interactions around residue mutations of interest were also investigated, by comparing them with their equivalent position in the M^{Pro} reference, using their modeled structure. These mutations comprised residues that underwent a higher number of mutations (G15D/S, V157I/L, and P184L/S) in addition to mutations that occurred at or close to the dimer interface (A7V and A116V). A7V significantly increased the number of

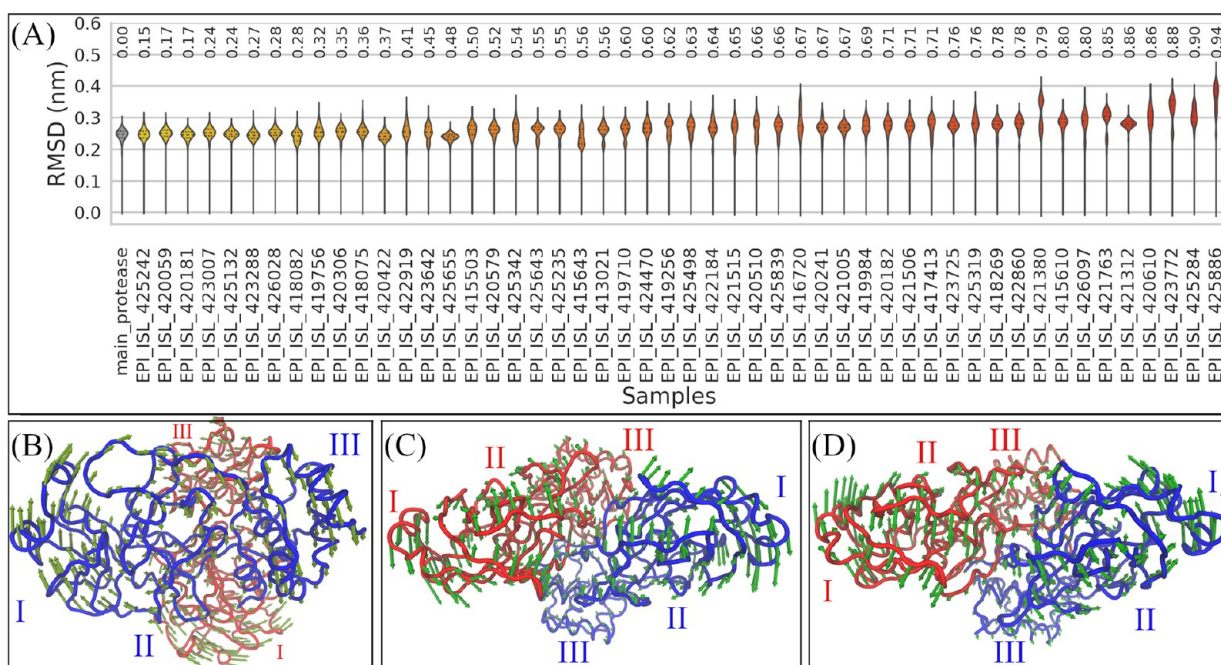


Figure 3. (A) Violin plots of C_{α} RMSD values for the reference (in gray) and the mutant (colored in blue) M^{PrO} , showing the 25th, 50th, and 75th percentiles in dotted lines inside the kernel density plots. Distributions are scaled by area and have been sorted by the KDE d_1 distance (shown above each distribution) computed between each sample and the reference protease. (B) General twisting motion of the protomers observed across M^{PrO} samples, inferred from the highest probability density reference pose using the ANM mode 7. (C) Mode 8 shown for the same reference pose and for (D) highest recorded distribution mode in EPI_ISL_425886 at 77 ns.

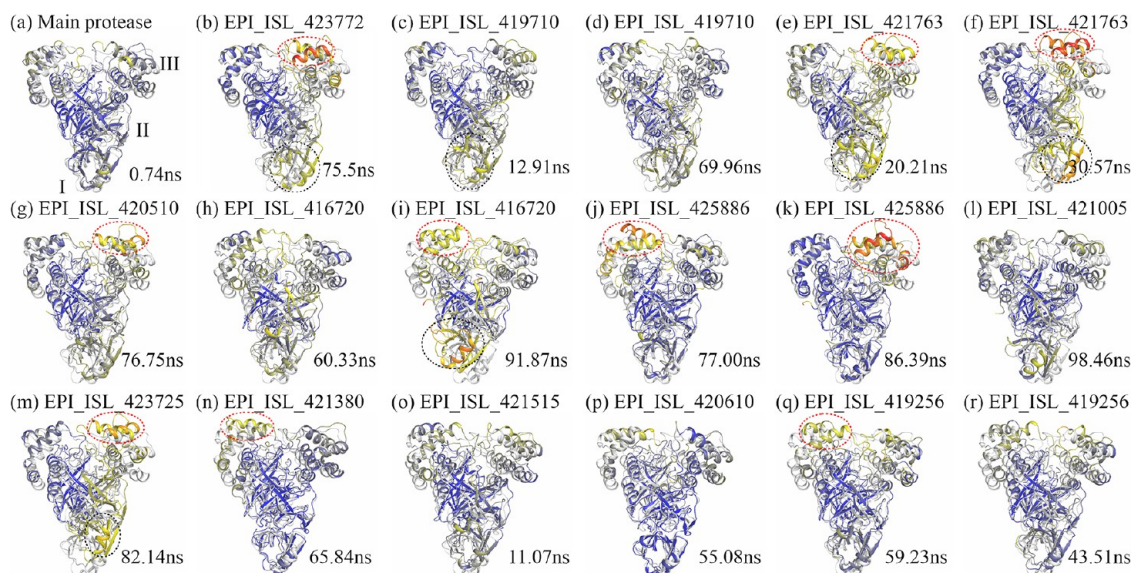


Figure 4. (a–r) 3D visualization of M^{PrO} conformations extracted from KDE distributions of C_{α} RMSD. All panels share a common color scale (ranging from blue through yellow to red) depicting the C_{α} distances calculated from an aligned high probability reference conformation obtained at 67.84 ns (in white cartoon representation). Regions of appreciably higher distances are circled in red (domain III) and black (domain I) dotted lines.

proximal interactions to neighboring residues when compared to the reference protein (Figure 2) and gained in hydrophobic interactions, although a clash in the van der Waals radius is additionally present.

G1SD is found to increase the number of proximal contacts by the largest extent while also modestly increasing the number of hydrophobic interactions. By replacing alanine with valine at position 116, an increased amount of proximal interactions is gained at V116 in sample EPI_ISL_425284, with a higher

amount of hydrophobic contacts to the residue. Mutations V157I/L both reduced the number of local hydrophobic contacts and resulted in a reduced van der Waals clash compared to the reference. P184L had a reduced number of proximal contacts compared to the reference, while P184S was very similar to its equivalent position in the reference. These observations only give a general indication of the local changes present in the static starting structures. In the sections that follow, the dynamic aspects of M^{PrO} are investigated.

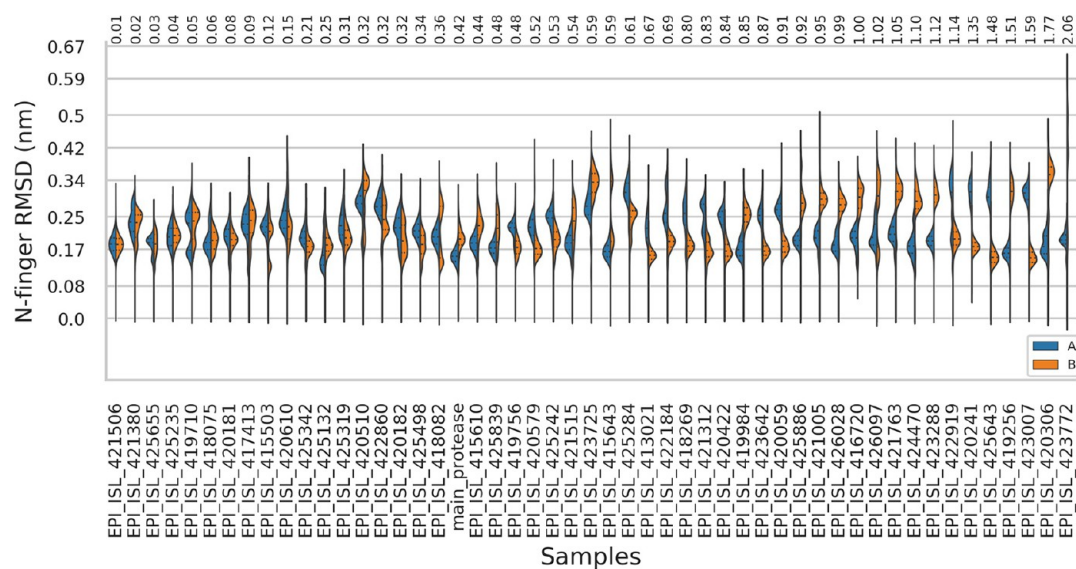


Figure 5. Kernel density distributions of RMSD values for the N-finger region across the mutant and reference protease complexes. The violin plots are split into two for each protein sample, showing the RMSD values for chains A (in blue) and B (in red). The tips of the distributions mark the minimum and maximum values for both chains combined in each complex. Samples have been sorted by increasing median distance between the chains, also shown (in Å) at the top of each sample distribution.

3.3. Estimation of the Protein Backbone Flexibility from MD Using C_{α} RMSD.

The C_{α} RMSD values obtained after frame fitting to the initial frames and periodic image correction (Figure 3A) revealed a relatively high range of divergence from the reference M^{Pro} among the 50 isolates, ranging from 15 to 94%. From the shapes of the kernel density plots, it can be seen that some mutants may equilibrate around single energy minima (for unimodal distributions), while others select for multiple major conformations, as depicted by the multimodal shapes. In all, these samples involve mutations A7V, M17I, A70T, A116V, K236R, Y237H, D248E, A266V, and N274D, in which the last five are exclusive to domain III. A7V occurs on the N-finger, which is a critical region for M^{Pro} dimer stability. M17I occurs on an internal loop that connects a β strand to a helix in domain I, while A70T occurs on a solvent-exposed loop in the same domain. Mutation A116V occurs in a buried β strand within domain II. From their visibly positively shifted upper quartiles (and based on the relatively large KDE d_j and the visibly higher modes), we may infer that more mobile backbones were sampled for mutants EPI_ISL_421380, EPI_ISL_423772, and EPI_ISL_425886 compared to the reference. Upon visualizing the trajectories, a slight, global twisting motion was observed between their protomers, whereby each protomer moved in opposite directions, while being tethered at the center. However, a similar motion was also generally observed in all other samples and is summarized using the first non-trivial mode (number 7) from the ANM using the highest probability equilibrium pose from the reference (Figure 3B). This may indicate that the global twisting motions are a normal behavior of dimeric M^{Pro} , at least under our simulated conditions for the apo state. On the other hand, the second non-trivial ANM mode (number 8) showed the domain rotational motions—they are similar overall for both the reference (Figure 3C) and in EPI_ISL_425886 where the highest equilibrium backbone RMSD was observed (Figure 3D), with the exception that protomer A behaved as protomer B, and vice versa in each case. The predictions were obtained from ProDy in VMD.⁹⁶

We further examined these samples by collecting single conformations found at each of the local peaks (local energy minima) from their individual KDE distributions. A cutoff KDE d_j value of 60% was used as a prior filter to limit the number of samples to be assessed visually. Using this criterion, modes from the KDE distributions of 25 mutant samples (marked by an asterisk symbol in Table 1), in addition to the reference protease, were used to extract conformations (Figure 4) corresponding to different local energy minima.

For a fair comparison, the reference protease conformation corresponding to the highest KDE peak was used as a comparator for all samples. Thereafter, conformations displaying the highest displacements were visually selected. From a bird's eye view of the sampled backbone RMSD in Figure 4, it can be seen that the regions of highest divergence include part of domain III and to a lesser extent domain I. Only one subunit is seen to display more movement in each of the extracted high probability density poses. Due to the homodimeric nature of the protein, a permutation of structural superpositions of the mutant samples was performed against the reference protease chains to determine the best reference position, based on the lowest RMSD to domain II, which was the least mobile protease domain during MD. Domain I also showed notable movement, albeit of lesser magnitude than domain III. This region mainly comprised the active site components. The fact that motion happens about a more stationary center is reminiscent of a fulcrum. The overall backbone dynamics and asymmetries are similar to what is described in SARS-CoV M^{Pro} .

3.4. Estimation of the N-Finger Flexibility from MD Using All-Atom RMSD. The N-finger region is an important structural foundation required for the stabilization of the functional M^{Pro} dimer, as reported by a large body of work done in SARS-CoV.⁹⁷ The N-finger from one protomer controls the activity of the dimer's alternate subunit.⁹⁸ Early work by Chou and coworkers demonstrated the importance of the salt bridge found between ARG4 (N-finger) and GLU290 (domain III) from alternate protomers in subunit associa-

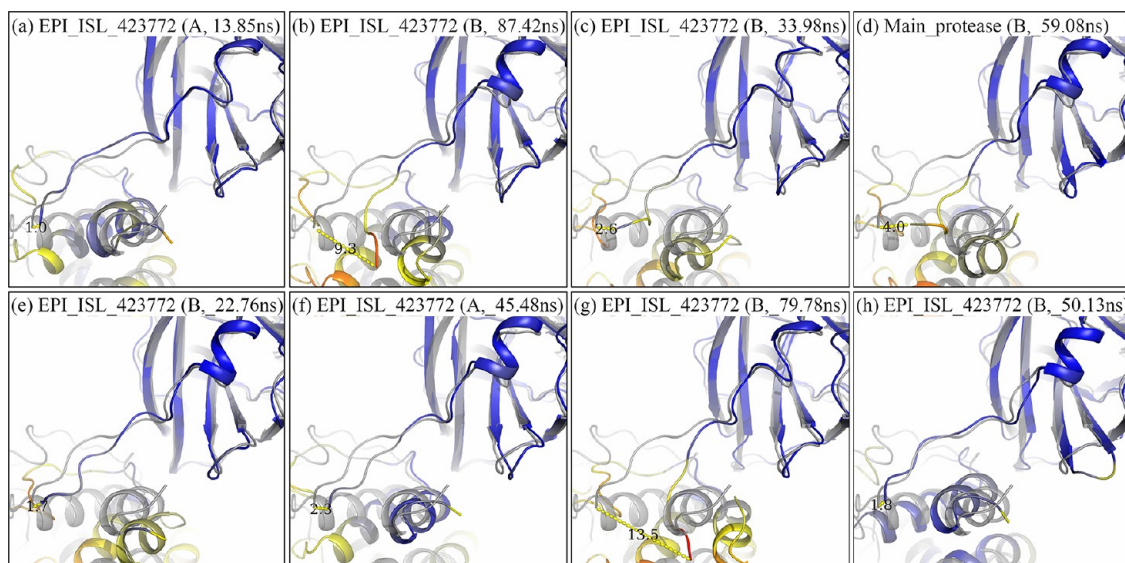


Figure 6. (a–h) 3D visualization of equilibrium N-finger conformations from different M^{Pro} samples. Each panel shows equilibrium poses of the N-finger using all-atom RMSD as the collective variable. The mutant chain labels are shown, next to the sampled times. The reference pose (chain A, 46.34 ns in gray) is used as a comparator in each panel, while the samples are colored according to their C_{α} atom displacements from the reference, starting with blue (lowest) through yellow to red (highest). The distance values are colored from blue (low values) through yellow to red (highest values). The distance between the first C_{α} atom (in Å) of each sample from the equivalent position in the reference protein is also shown.

tion.⁹⁹ Additional studies involving deletion⁵³ and missense mutations^{51,100} have highlighted the importance of the coronaviral M^{Pro} N-finger region in enzymatic activity. Multiple studies have also reported the dimer to be an active form of the enzyme,^{51,100} with only one protomer being active at a time.⁵² Chen and colleagues suggested that the N-terminus only stabilizes the active form of the enzyme and is not a prerequisite for dimerization¹⁰¹ of the SARS-CoV M^{Pro} . More specifically, an N-finger from each protomer interacts with the GLU166 residue from the alternate protomer to stabilize the catalytic pocket.¹⁰² More recently, Li and coworkers showed that enzymatic activity is in fact coupled to the flexibility of a short loop (residues 139–141) near the active site by engineering an active tetramutant M^{Pro} monomer.¹⁰³ Their work, which was focused on understanding the stability of SARS-CoV M^{Pro} , obviates the need for a similar inquisition in its SARS-CoV-2 homolog. Therefore, we started our investigation of the N-finger stability within the dimeric protein across all M^{Pro} samples, as a means to detect any hint of reduced enzymatic activity via a reduction of dimer stability.

While the observation of dissociation would give a clear verdict about inactivation (as experimentally reported in SARS-CoV^{98,100}), its absence during an MD simulation may not inform us about inactivation. Independent simulations performed in duplicate (not shown) of dimeric M^{Pro} containing the mutants G11A, E14A, and R298A and even of the triple mutant (G11A/R298A/Q299A) did not dissociate over a period of 100 ns, even though these were experimentally shown to be inactive monomers in SARS-CoV, suggesting that particular equilibrium states may need to be traversed before observing such events. Another possibility may lie behind an incompletely understood mechanism of dimer formation, for which two models have been proposed using SARS-CoV, both suggesting that M^{Pro} dimerization may be substrate-induced.^{102,104} We nevertheless examined the 100 ns dynamics for any notable differences between the SARS-CoV-2 variants.

After fitting the proteins globally, no further fitting was done for the N-finger RMSD calculation in order to better represent the N-finger motions. As seen in the KDE plots (Figure 5), the reference protein has two very stable N-finger conformations, characterized by unimodal distributions although the protomer equilibria are asymmetric. This tendency toward asymmetry is seen in most of the cases, with the exception of samples EPI_ISL_421506, EPI_ISL_417413, and to some extent EPI_ISL_425235. Due to the existence of both forms of symmetry, with asymmetry largely dominating the dynamics, it is probable that some form of dynamically associated constraint may drive one N-finger into an equilibrium state while the other is forced into another. While one may speculate that EPI_ISL_421506 may be in an inactive state due to the perceived symmetry, other parts of the enzyme are not asymmetric. Multimodal distributions were observed in several cases, which clearly suggest the presence of multiple local minima, with transitory states occurring in between these modes. On the far right of Figure 5, samples EPI_ISL_423288, EPI_ISL_422191, EPI_ISL_420241, EPI_ISL_425643, EPI_ISL_419256, EPI_ISL_423007, and EPI_ISL_420306 show median differences above 1.1 Å between the protomers. EPI_ISL_423772 showed the most prominent difference, which corresponds to a larger amount of time spent away from a stable conformational equilibrium, even though a stable equilibrium was also visited in chain A (the lowest mode for the same sample).

N-finger poses corresponding to the RMSD KDE peaks for EPI_ISL_423772 (M171 mutant) and the reference protease are shown in Figure 6. From the N-finger visualization, we can observe that its distance from the high probability reference protease equilibrium pose varies from 1.0 to 13.5 Å. While a subfigure (g) shows the highest distance reached by the N-finger C_{α} atom, all other poses (a–f and h) in Figure 6 correspond to multiple modes from the KDE distribution. Additionally, from the chain labels, the asymmetry can also be observed. It is also noteworthy to observe the transient nature

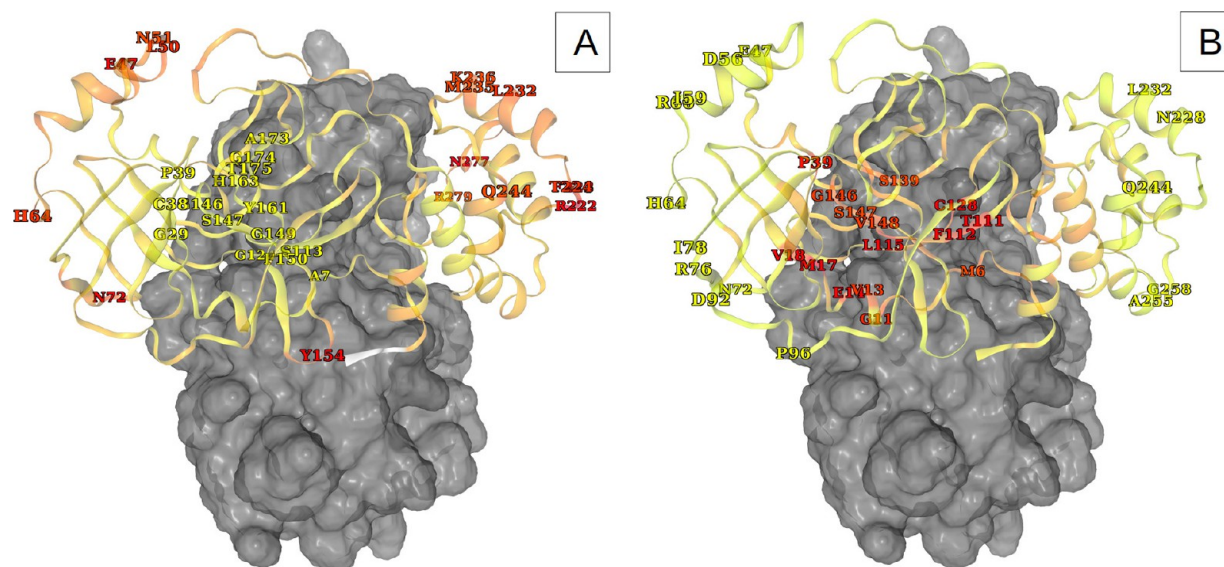


Figure 7. 3D mapping of averaged values for (A) RMSF and (B) average BC, computed across all M^{Pro} samples. Only the extremes (top and bottom 5% of the ECD values across all samples) of averaged values are labeled for each metric. The lowest averaged values are in yellow, while the highest ones are in red. The last three C-terminus residues (in white) were not mapped by RMSF as their high values would mask other values. While only one protomer is detailed, the data is applicable to both protomers. The other protomer is represented as a gray surface.

of the α -helical region (residues 10–15) to which the N-finger is connected, both in the mutant and the reference. The STRIDE¹⁰⁵ functionality within VMD (version 1.9.3) predicts the helical region to be shorter (residues 14–16) in the reference protease. This lengthening and shortening are likely related to interprotomer stability and may potentially be associated with enzymatic activity. It may therefore be a generalization of a similar functionality observed in the chameleon loop that is directly associated with enzymatic activity switching. The same protein displayed discernible backbone motions in domains I and III (Figures 3 and 4). While the mutation occurs on a beta strand located in a core area of the protein, the non-bonded interactions are similar for both M17 and I17.

It is clear that the N-finger can adopt a range of equilibrium conformations, with some showing protomer symmetries and others not. Above all, it can safely be assumed that all samples are enzymatically active; otherwise, the SARS-CoV-2 would not be able to perform catalysis and thrive in the human host. The distribution shapes may therefore be different parts of a continuum along the potential energy surface of active M^{Pro} . It is also possible that a different energy surface may manifest itself in the presence of a substrate to emphasize the differences in distribution shapes. There are hypotheses that support this claim, and they are elaborated further in section 3.8.

3.5. Analysis of Residue Fluctuations and BC across M^{Pro} Samples. The analysis of residue fluctuations shows that there are generally interspersed areas of low and high flexibility along M^{Pro} (Supporting Information, Figure S1). Some local areas of higher flexibility were also seen, as is generally observed in unstructured secondary structures.¹⁰⁶ Additionally, from the lack of clustering between chains belonging to the same sample for each segment of M^{Pro} (Supporting Information, Figure S1; a cluster tree was removed for figure visibility), we conclude that the residue dynamics for the same domain between alternate chains of M^{Pro} are asymmetric in the dimeric apo form. Focusing onto the specific regions of M^{Pro} , it can be seen that the N-terminal region of the N-finger

generally displayed moderate residue fluctuations despite being sandwiched at the interface of two M^{Pro} chains, thus agreeing with the N-finger RMSD data. Domain I is generally moderately flexible at residue positions 22–25, 33, 34, 92, and 93. Higher flexibility was globally observed along the intervals 45–65 and 71–74 and at position 76. Very high fluctuations were recorded for a subset of mutants at positions 46–54, most particularly for only one chain among the mutants EPI_ISL_416720, EPI_ISL_423642, EPI_ISL_415503, and EPI_ISL_425886, thus reinforcing the observation of asymmetry between chains. Mutations from samples EPI_ISL_416720 and EPI_ISL_425886 were already found to lead to increased backbone fluctuation using C_{α} RMSD. The Y237H mutation in EPI_ISL_416720 introduces two carbon H-bonds (one with L272 and another with V233), in addition to the pi-alkyl interaction that is present in both the reference and this mutant, which seem to hold the solvent-exposed helices together in domain III. In a review by Horowitz and Trievel, the carbon H-bond was highlighted as an underappreciated interaction that is otherwise widespread in proteins, with the ability to form interactions as strong as conventional H-bonds via polarization.¹⁰⁷ It is possible that such an increase in interaction may improve the stability around this region in domain III for the Y237H mutation. In the case of the D248E mutation in EPI_ISL_425886, the D248 side chain is H-bonded to Q244 in the reference structure, possibly stabilizing the helical structure. This interaction is absent upon mutating to an E248. In the last frame from MD, the side-chain epsilon oxygen was found to interact with its backbone hydrogen atom, indicating a decreased stabilization of the helical region of domain III. T201A in EPI_ISL_423642 abolishes the H-bond that is otherwise present between T201 and the backbone oxygen atom of E240, very likely weakening their interaction. The V157I mutation in EPI_ISL_415503 does not significantly alter the non-bonded interactions but occurs on a beta strand on domain II. A unique behavior was observed in the case of chain B of EPI_ISL_423772, where the leading residues of the

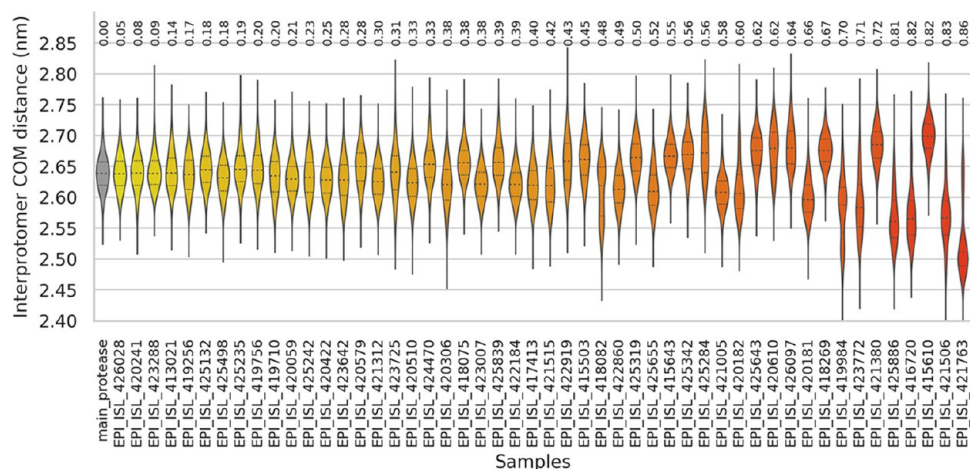


Figure 8. Distributions of interprotomer COM distances across samples, arranged in an ascending order of the KDE d_j . The reference protease is in gray, while the mutants are colored from yellow to red in an increasing order of distance from the reference KDE.

N-finger were the most flexible while the rest of the protein was the least flexible across all samples. However, the M171 mutation does not significantly change the non-bonded interaction in EPI_ISL_423772 but occurs on a beta strand in domain I. Domain II is most flexible across all samples at residue position 153–155. Moderate flexibility is generally observed at residue positions 100, 107, 119, 137, 141, 142, 165–171, 178, and 180. The linker region was generally highly flexible in all cases within the region spanned by residues 188–197. Notably higher fluctuations were observed at position 185 in samples EPI_ISL_420610 (chain B) and EPI_ISL_423007 (chain A). Across all domains, parts of domain III contain the most flexible residues within M^{Pro}. It is highly flexible at residue positions 222–224, 231–233, 235, 236, 244, 245, and 273–279. The high fluctuations observed at C-terminus residues may not be very informative in our case as they freely interact with the solvent and do not have a strong enough network of non-bonded interactions with the M^{Pro} domains. As a whole, from the empirical cumulative distribution (ECD) of averaged root-mean-square fluctuation (RMSF) values across all M^{Pro} samples, the top 5% positions (most variable regions) comprise residues 222, 277, 223, 154, 47, 72, 50, 224, 232, 64, 279, 236, 235, 244, and 51, with values ranging from 0.386 to 0.245 nm. The bottom 5% (most stable regions) of the distribution comprise positions 149, 146, 147, 174, 29, 113, 163, 39, 124, 150, 7, 175, 38, 161, and 173, with RMSF values ranging from 0.057 to 0.077 nm. On the 3D mapping (Figure 7A), it can be seen that the regions with the highest flexibility are solvent-exposed surfaces, comprising loops or parts of helices connected by loops. The central core of the enzyme has the lowest flexibility, most likely to provide structural stability to the functional dimer. Catalytic residues (HIS41 and CYS145) are connected to these stable core residues on one side. However, HIS41 is connected on the other side to a more mobile structure composed of a 3_{10} helix connected by loops on each end, which forms a lid structure, similar to what is described earlier for previous human coronavirus strains.^{108,109} Further discussions around the substrate binding site are given in section 3.8.

BC is a network centrality metric that is maximized when a large number of nodes (C_β or GLY C_α atoms) traverse a single node along geodesic paths to reach a large number of other nodes within a network. When averaged from MD frames, this

metric has been shown to be approximately inversely related to RMSF.¹¹⁰ These values were very similar across all samples (Supporting Information, Figure S2). For this reason, the discussion of our findings will be around the overall BC behavior recorded across all samples. High and low BC values are present within all domains (Supporting Information, Figure S2). The N-finger was found to be generally composed of high BC residues, most likely due to its high degree of non-bonded interactions between the protease chains. A large portion of domains I and III have low BC values, most likely due to the comparatively lower amount of contact with protein surfaces, which allows for a relatively higher mobility compared to domain II. The top 5% highest averaged BC values across all M^{Pro} samples were either found in the monomer core regions or at the dimer interface, as seen in Figure 7B. These comprised residue positions 17, 128, 115, 111, 112, 14, 18, 39, 11, 13, 146, 148, 147, 139, 6, and 140 in a descending order of overall averaged BC values, ranging from 10480.987 to 6064.650. The residue positions within the lowest 5% overall averaged BC values comprise the positions 96, 72, 92, 258, 64, 244, 47, 59, 228, 56, 76, 60, 255, 78, and 232, ranging from 8.095 to 46.251. A complete listing of the top BC residues for each protomer is given in Table S2. Compared to the high BC areas that correspond to very well maintained communication residues (which may well be actual functional sites¹¹⁰), the low BC values represent residues that are least important for maintaining the flow of communication across the protein due to the transient nature of their medium- to long-ranged path lengths. From the computed sample average BC values and as seen in the Supporting Information, Table S2, residue positions 17 and 128 were found to occur as the most common first two residues in all cases. In the previous work done in human heat shock protein, high BC residues found within cavities were found to correspond to allosteric hot spots, as these had been independently verified by the sequential application of external forces on protein residues using the perturbation response scanning (PRS) method. In our case, however, these two positions are not found in cavities and are rather buried structural units that are not very mobile in the short to medium range. The high BC at M17 is possibly due to the increased stability imparted by their presence at a dimer interface. Due to its presence between the N-finger and the beginning of domain I, it is likely a pivot point for the N-

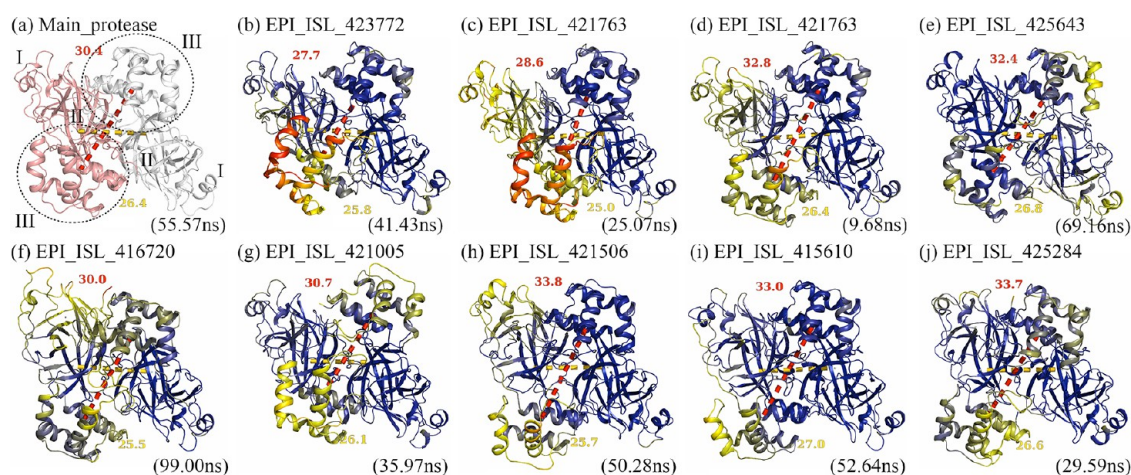


Figure 9. (a–j) 3D visualization (top view) of the interprotomer COM distances across samples. Domain III from each protomer is circled in black, while the chains are colored white and salmon in the reference (a). All other samples are colored according to their C_{α} atom displacements from the aligned reference, starting with blue (lowest) through yellow to red (highest). The interprotomer COM distance is depicted by horizontal yellow dashes (and yellow font). Domain III COM distance is also shown, in red dashes and font.

finger. Due to the high centrality of these residues, it is possible that mutations leading to the alteration of their physicochemical activity may be accompanied by a decrease in dimer stability. EPI_ISL_423772, which contains the M17I mutation displayed a relatively high backbone mobility, more particularly along one of the α -helices of domain III. Even though both residues are hydrophobic, an increase in backbone motion was observed. It is possible that by the replacing of the unbranched methionine side chain¹¹¹ by the branched amino acid isoleucine, the potential for local mobility is reduced, possibly leading to far-ranging compensatory effects.

3.6. Estimation of Interprotomer Distances Using COM Distance. The COM distance distributions estimated from all atoms between the M^{Pro} protomers (Figure 8) indicate that the distance between them can vary from over about 2.4 to 2.8 nm globally, and the presence of modes of differing height suggests at least two different equilibrium chain COM distances in these cases. About half of the samples displayed KDE d_j distances that were very close to that of the reference protease, corresponding to a distance value of about 0.5 nm. More specifically, in isolates EPI_ISL_421763 and EPI_ISL_419984, a significant proportion of the sampled conformations depicts the presence of closer chain COM values, though the percentage of such conformations is lower in EPI_ISL_419984. 3D visualization showed that the THR70 side chain formed an additional (but transient) H-bond with either VAL73 or with the MET17 backbone oxygen in EPI_ISL_421763. In addition to increasing interconnectivity within the β hairpin, temporary H-bonding interactions between MET17 and THR70 connecting the N and C-termini from domain I may be the main reason for the lowest interprotomer COM distance. We note that ALA70 cannot form such a bridge with MET17 due to its hydrophobic side chain. From Figure 9, it can be seen that domain III is a key player associated with an increased COM distance, given that the largest displacements are experienced there. Therefore, it is not impossible that THR70 may have remotely interfered with the motion of domain III. Longer MD simulations may be needed to ascertain this observation. Isolate EPI_ISL_415610 (bearing mutation A193V) displayed the highest median interprotomer COM distance. The main physicochemical difference in A193V is the higher molecular weight and

volume of the valine side chain; however, the mutation occurs in the solvent-exposed linker that connects domains II and III. Weighted residue contact mapping showed that the main difference was a six-fold increase in the residue contact frequency (0.72 in the mutant versus 0.17 in the reference) between residue 193 and THR196, in chain A. Samples starting from EPI_ISL_425655 onward were sampled for modes, aligned to the reference protease, and colored by their C_{α} – C_{α} distance in Figure 9. Only samples showing the most marked differences from the reference are shown. Superposition adjustment was made as previously explained to obtain the most probable structural alignment for the dimer. Due to the observation of a generally higher deviation of domain III from the rest of the protein, a correlation was calculated at a 95% significance level for all short-listed equilibrium conformations against the interprotomer COM distance, and as intuited, a non-negligible correlation of 0.59 with a p -value of 0.003 was obtained. It is very likely that the domain III COM motions could drive larger conformational changes due to their size and flexibility, should their stabilizing interprotomer interactions weaken. This is in good agreement with previous studies suggesting its involvement in regulating of dimerization, more specifically with the proposition by Hu and colleagues that domain III might act as a “motivator” controlling N-finger movement in SARS-CoV.¹¹² Another notable observation is the relatively higher mobility of one of the domain III α -helices and one of the 3_{10} helices in some mutant equilibrium conformations, when compared to the homologous positions in the reference equilibrium conformation. COM distance can be influenced to a certain extent by the shape of domain arrangements as it can modulate the COM. As such, this may not represent other changes to the protein geometry. Therefore, we proceeded to quantify interdomain angles as a next step.

3.7. Estimation of Interdomain Angles in Each Protomer of M^{Pro} . The angle formed between domains I, II, and II has a relatively high information content for M^{Pro} dynamics due to the variability that it captures. While both chains A and B behave similarly in the reference, a higher degree of interdomain angle variation is seen across the mutants, comprising skewed distributions and angle asymmetries between protomers (Figure 10). The angle distributions

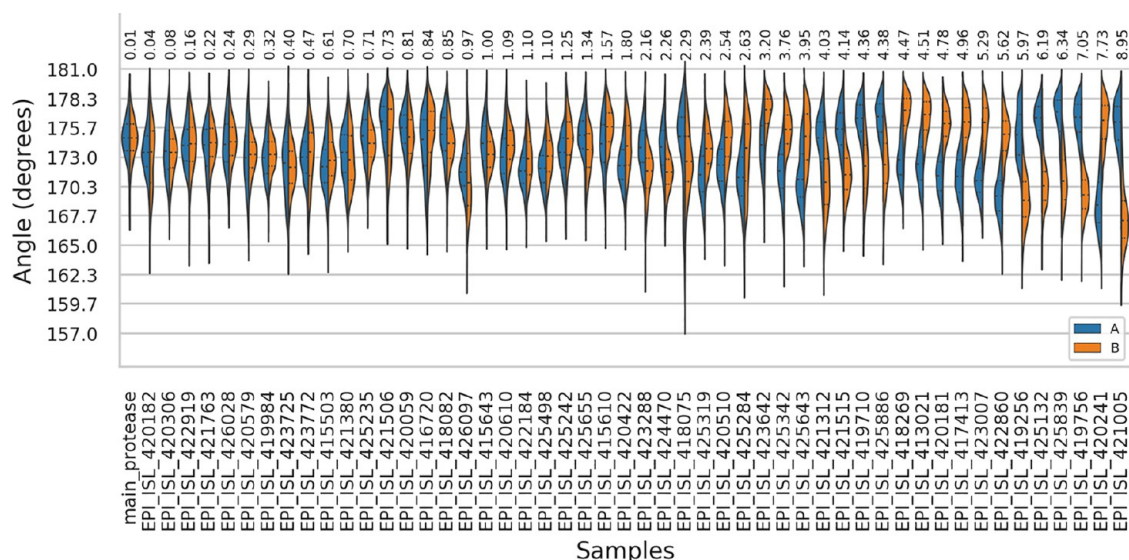


Figure 10. Kernel density distributions of interdomain angles (domains I–II–III) across the mutant and reference protease complexes. The violin plots are split into two for each protein, showing the interdomain angles for chains A (in blue) and B (in red). The tips of the distributions mark the minimum and maximum values for both chains combined in each protein complex.

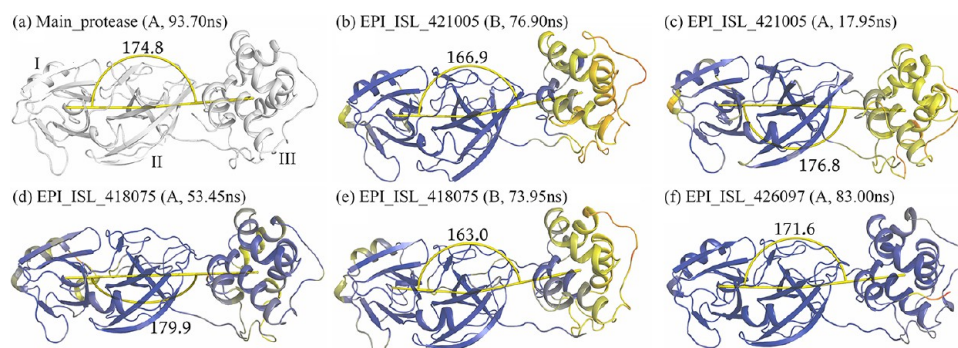


Figure 11. (a–f) Visualization of interdomain angles (domains I–II–III) across the mutant and reference protease complexes. The reference is in white, while the mutant samples are colored according to their C_{α} atom displacements from the reference protease, starting with blue (lowest) through yellow to red (highest). Angles are shown in degrees. Only one of the protomers is shown in each case.

are similar to those of the reference in some cases but are also more variable in others. The difference between interdomain angles was computed for each protomer pair and used to sort the samples in an increasing order of divergence, such that the most asymmetric distributions occur on the right. Among the divergent samples, using a median difference of 2 degrees as cutoff, chains A and B were found to be most divergent in samples EPI_ISL_423280, EPI_ISL_424470, EPI_ISL_418075, EPI_ISL_425319, EPI_ISL_420510, EPI_ISL_425284, EPI_ISL_423642, EPI_ISL_425342, EPI_ISL_425643, EPI_ISL_421312, EPI_ISL_421515, EPI_ISL_419710, EPI_ISL_425886, EPI_ISL_418269, EPI_ISL_413021, EPI_ISL_420181, EPI_ISL_417413, EPI_ISL_423007, EPI_ISL_422860, EPI_ISL_419256, EPI_ISL_425132, EPI_ISL_425839, EPI_ISL_419756, EPI_ISL_420241, and EPI_ISL_421005 in an ascending order of difference. This provides additional support behind our general observation of the protomers generally behaving in an independent manner in SARS-CoV-2 M^{Pro} , as similarly described in SARS-CoV. As explained earlier, symmetric behavior may be part of the protease mechanics in the absence of a substrate. Due to the richness of information retrieved from the measurement of interdomain angles, we propose that

this metric may also help assist the *in silico* characterization (or differentiation) of M^{Pro} variants from future strains of SARS-CoV-2 should any particular virus-associated phenotype becomes available.

Samples of the most divergent conformations from the reference protease are shown in Figure 11. Using domain II as a fulcrum, it can be seen that interdomain angles are not necessarily planar. On the basis of divergence from the reference protease, it can be observed that domain III moves more than domain I around the fulcrum. While the equilibrium for EPI_ISL_418075 is quasi-planar (even though it displayed the widest range of interdomain angles), the equilibrium displaying the largest angle shown (183.2 degrees) was recorded from EPI_ISL_421005 (from the exponent of the shown angle in Figure 11c).

3.8. Analysis of Pocket Compaction and Proposition of a Dual Allosteric Site in SARS-CoV-2. The substrate binding pocket, nested in the cleft between domains I and II,¹¹² is composed of the very flexible loops intertwined with the catalytic dyad residues HIS41 and CYS145. With the assistance of HIS41, residue CYS45 acts as an electrophile during the first step of hydrolysis.³⁶ Previously, studies in SARS-CoV have reported that components of this pocket—the

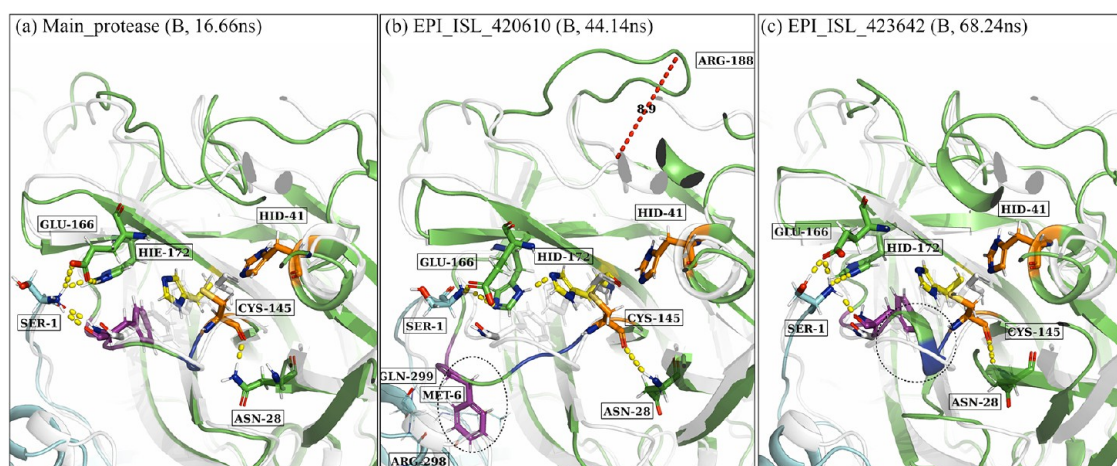


Figure 14. 3D visualization of the M^{Pro} substrate binding site poses of interest. (a) Local minimum reference from the reference protease conformation. (b) Detection of a non-canonical pose for residue 140 (in purple) together with a significant loop opening (dashed red line) in EPI_ISL_420610. (c) Early stages of a 3_{10} helix formation in EPI_ISL_423642. Another local minimum (highest probability density) pose from the reference protease (in white) is used as a comparator in all panels. Chain A is in pale blue, while chain B is in green. The key interacting residue for PHE140–HIS163 is depicted as yellow sticks.

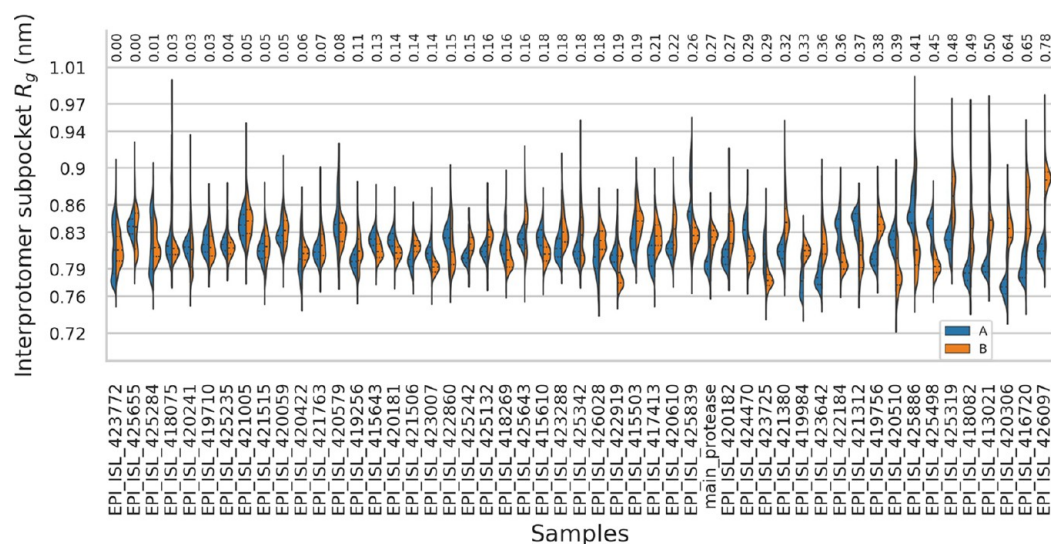


Figure 15. Kernel density distributions of R_g values across samples for the mirrored interfacial (and potentially allosteric) pockets. The differences, shown at the top, are in nanometers.

could play an important role in the dimerization properties of the protomers. The substrate binding site from each protomer was also examined due to its already known functional importance in catalysis. While the substrate binding pockets are easily defined as belonging to a given protomer, the interfacial pocket is composed of residues from each chain, namely, residues 116, 118, 123, 124, 139, and 141 on chain A and residues 5–8, 111, 127, 291, 295, 298, 299, 302, and 303 on chain B. The mirrored interfacial pocket comprised the same residue positions but for the opposite chain labels. When compared to available literature for the SARS-CoV, multiple residues corresponded to stabilizing elements of the catalytic pocket. For instance, residues 139 and 141 form part of the “chameleon” switch, which assumes a 3_{10} helix form when inactive, while residues 5–8 are part of the N-finger, which is essential for the stability and activity of the alternate protomer. The FTMap probes formed identical cross-clusters composed of probe IDs 1amn, 1ady, 1eth, and 1acn at each site, which respectively correspond to the compounds methylamine,

acetaldehyde, ethane, and acetonitrile. As noted from the probe chemical compositions, this dual site has the potential to accommodate compounds of low molecular weight, with no (or probably limited) rotatable bonds and low octanol–water partition coefficients.¹¹⁴ The substrate binding site, as determined by PyVOL, comprised residues 25, 26, 27, 41, 44, 49, 140, 141, 142, 143, 144, 145, 163, and 166 in each protomer.

From the distributions of R_g values for the catalytic site (Figure 13), we promptly observe the asymmetry in compaction between substrate binding pockets coming from each protomer in each sample. Partial symmetry is seen in few cases, such as EPI_ISL_419710, EPI_ISL_425242, EPI_ISL_423007, 416720, EPI_ISL_421380, EPI_ISL_425284, and EPI_ISL_419256. In the reference sample, we observed a shift in equilibrium R_g , where one protomer oscillates around a value while its counterpart also explores two others. Multimodal distributions indicate the presence of more than one equilibrium, which hints at cavity expansion and

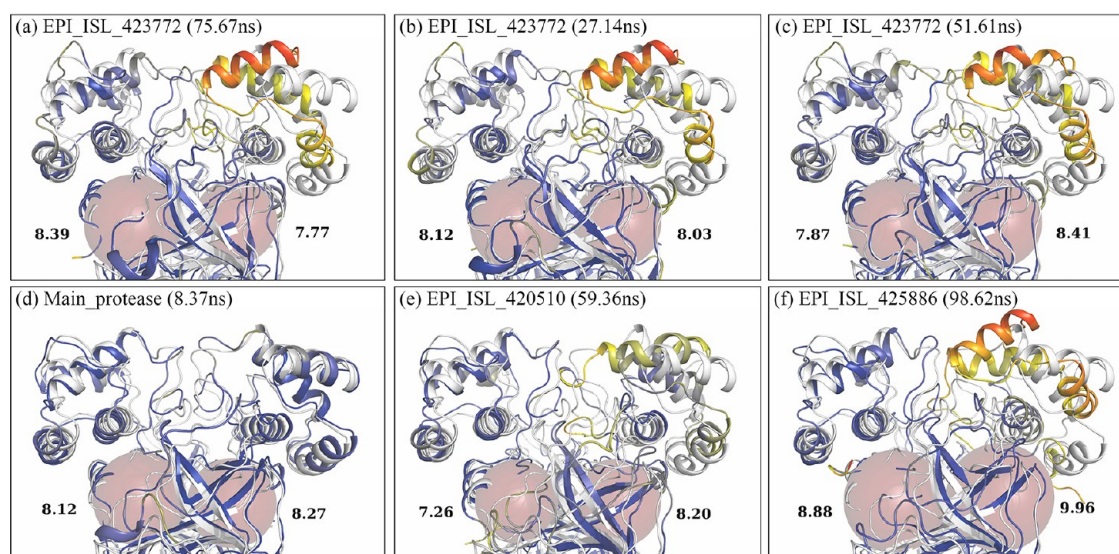


Figure 16. (a–f) Visualizations of the R_g values (Å) of the mirrored interfacial pockets in various samples. The reference (obtained after 63 ns) is in white, while the mutant samples are colored according to their C_α atom displacements from the reference protease, starting with blue (lowest) through yellow to red (highest). Regions of high divergence are circled. The spheres are centered at the COM of the pocket residues, while their radii are the R_g values in Å.

compaction movements. The most compact substrate binding cavities are observed in one of the protomers from sample EPI_ISL_425489 and to some extent samples EPI_ISL_418082 and EPI_ISL_423642.

Some substrate binding cavities of interest are shown in Figure 14. Shi and colleagues described several features surrounding the substrate binding pocket in SARS-CoV.¹⁰⁰ One of them is the requirement of a key interaction between residues PHE140 and HIS163 for the stability and activity of the protease. This same region is shown in Figure 14, using poses identified from local energy minima for (a) the reference protease, (b) EPI_ISL_420610, and (c) EPI_ISL_423642. The highest probability density pose (in white) is used as a comparator. In Figure 14a, the stabilizing PHE140 and HIS163 display their stacked rings and the stabilizing H-bonds linking the N-finger SER1 to GLU166 and HIE172. ASN28 is also shown to stabilize the positioning of CYS145 in all cases. In EPI_ISL_420610 (N274D), however, we observed a yet undescribed relatively stable pose for PHE140. It basically loses its described interaction and flips to hydrophobically interact with the C-terminal helix, via ARG298 (pi-alkyl interaction), GLN299 (amide-pi stacked interaction and a carbon H-bond), and MET6 (pi-sulfur interaction).⁶⁵ This has the effect of pulling the oxyanion loop to face away from the catalytic residues, thus making the pocket wider. Also shown is the coupled movement of an opposite loop (containing ARG188) by a distance of 8.9 Å, compared to the reference protease conformation. In EPI_ISL_423642, the oxyanion loop showed signs of stabilization, which may correlate with a lack of enzymatic activity for that protomer. As this was observed for only one of the protomers and given that the isolate was obtained from an infected individual, for this sequence, we cannot infer global inactivation.

As done for the substrate binding cavity, the degree of compaction of the interprotomer cavities was also measured. Here as well, the distributions are asymmetric (Figure 15). It is tempting to do a comparison between the substrate binding cavity and the interprotomer pockets. However, when the median R_g values obtained for the substrate binding pocket are

correlated (using Spearman's correlation) against the corresponding medians for the interprotomer pocket across all samples, no significant correlation was obtained, even though in our case, residue 141 was shared between both pockets. On an individual level, however, there is a significant degree of correlation between the interprotomer pocket and the substrate binding site, ranging from -0.68 to 0.54 (using Spearman's correlation), with p -values < 0.01 and absolute correlations > 0.1 for 64.9% of the sample comparisons (two substrate binding sites vs. two interprotomer pockets for each sample). As the interprotomer pocket is mirrored between chains, this may explain the negative correlations. From this finding, we propose that the interprotomer pockets may play an important role in affecting the degree of compaction of the binding cavity and vice versa. This suggests the possibility of a potentially bivalent modulation of the interprotomer pocket and the substrate binding site. As R_g only captures the overall degree of compaction, it may not entirely inform us about the cavity volume accessible to an allosteric modulator; however, this may be an interesting lead for allosteric modulator targeting. Snapshots where large deviations were obtained (compared to a high probability density reference conformation) are shown in Figure 16. The R_g values are used as radii for the spheres shown in Figure 16, centered at the COM of the allosteric pockets. Different distribution modes for EPI_ISL_423772 (a–c) point to a greater movement of an α -helical region of domain III, from only one protomer at a time, as seen previously. The equilibrium from EPI_ISL_425886 and to some extent that of EPI_ISL_420510 behaved similarly. The second equilibrium from the reference sample (d) performed very similarly to the reference of highest probability density used as a comparator (in white) for all the shown cases. From all the equilibrium poses, it can be seen that the differences in subpocket R_g between each protomer pair are not associated with similar displacements of the domain III helix.

3.9. Investigating Correlated Residue Motions Using DCC. In order to compare the DCC across all samples, the DCC matrices were linearized before calculating pairwise

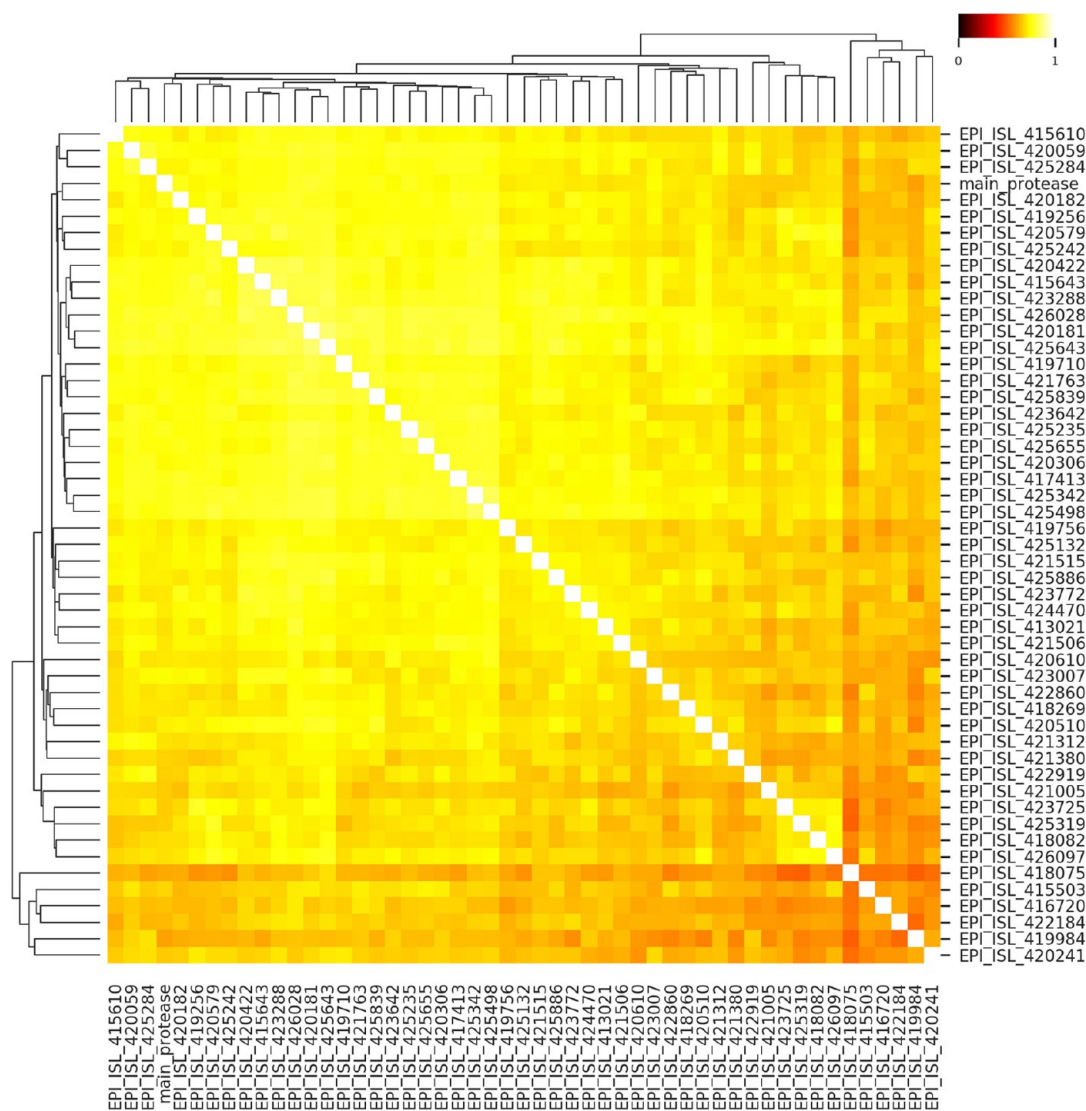


Figure 17. Heat map of correlations obtained from the linearized DCC matrices for the mutants and the reference M^{pro} clustered using the Euclidean distance.

correlations and clustering the final matrix (Figure 17). While abstracting out the intricate details of residue correlation, clusters of correlated samples inform us on the global similarity of correlated protein motion across each pair of samples. From Figure 17, it can be seen that the samples are generally highly correlated, with the exception of samples EPI_ISL_415503, EPI_ISL_416720, EPI_ISL_419984, EPI_ISL_420241, EPI_ISL_422184, and EPI_ISL_418075, which form a subcluster of moderately correlated samples. Both EPI_ISL_415503 and EPI_ISL_416720 have been described in section 3.5 as having higher RMSF values at positions 46–54 due to the mutations V157I and Y237H. Mutation R105H (occurring on a loop region) in EPI_ISL_419984 leads to the loss of an H-bond with F181 but forms a pi-pi stacking interaction with Y182. The most likely explanation for a difference in this case may be related to the sampling of at least two conformational equilibria using protomer COM distance, as described in section 3.5. In sample EPI_ISL_420241, the P184L mutation occurs in a solvent-exposed loop, and no major non-bonded interactions were detected from the side chains or backbone atoms. The main reason for the observed difference is due to the increased divergence in interdomain angles sampled from

MD. In the case of EPI_ISL_422184, as explained in section 3.7, it was found that the N-finger from one chain had moved by a larger extent at the end of the simulation (at about 93 ns), diminishing its contacts with the alternate protomer, to interact more with its own protomer. This may be attributable to the S301L mutation, which reduces H-bonding at the end of the C-terminal helical structure. In the reference, four H-bonds are formed between S301 and the backbone oxygen atoms of V297 and R298, whereas two H-bonds are formed by L301. Sample EPI_ISL_418075 had the lowest positive correlation to all samples. Upon closer examination, it was found that the terminal α -helix for each protomer was getting gradually destabilized toward the end of the simulation. This is very likely an artifact linked to the absence of the two C-terminal residues, as the residue is a solvent-exposed protein and displays very similar residue interactions at position 255 in both the reference and the mutant, even though A255V occurs on a helix.

The main observation across all samples is that residues within their individual domains are highly positively correlated within their respective domains. Domains I and II are seen to share a high degree of correlation with each other, behaving as

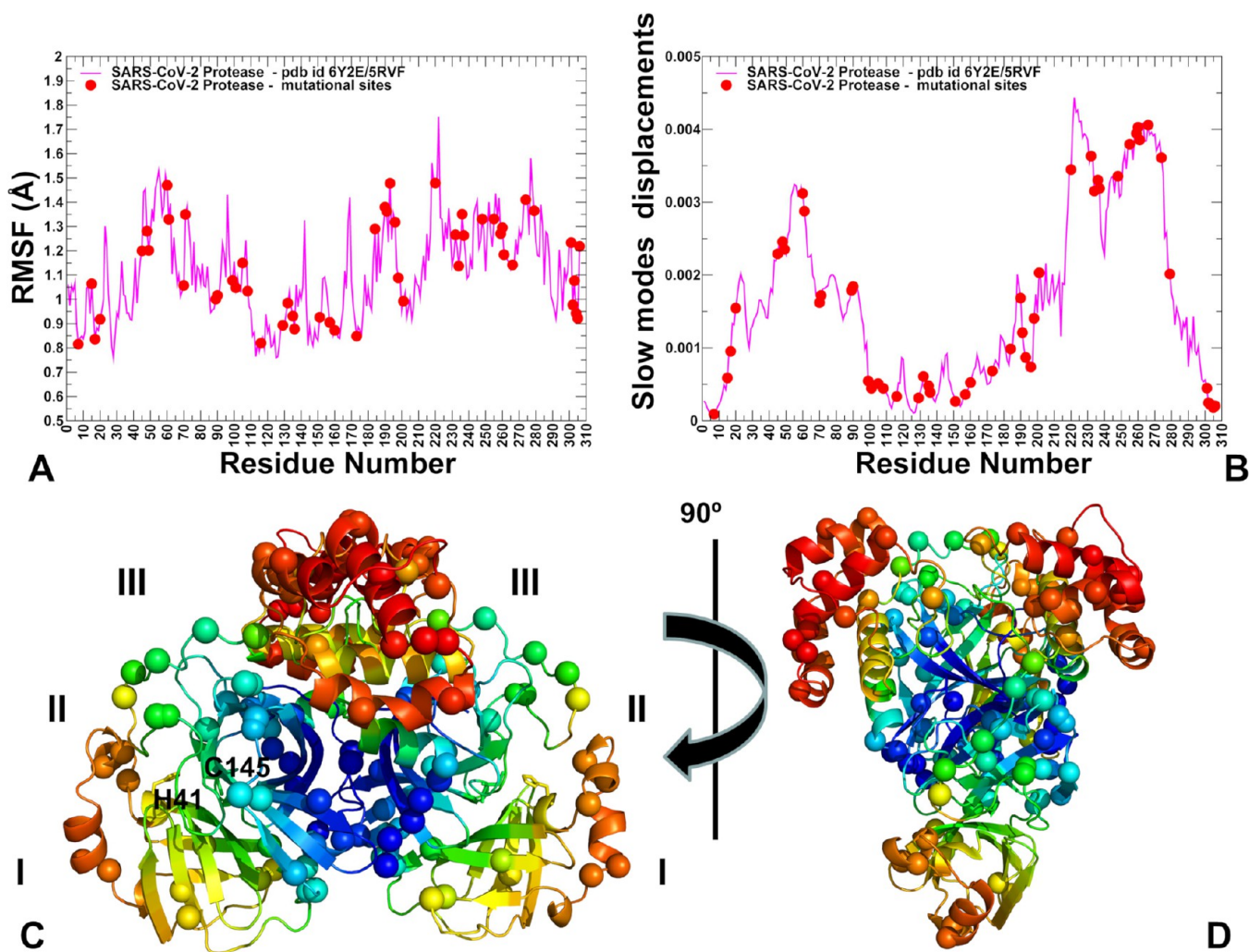


Figure 18. Conformational dynamics and collective motion slow-mode profiles of the SARS-CoV-2 main protease structures. (A) Computed root-mean-square fluctuations (RMSF) from CG MC dynamics simulations of the free enzyme of the SARS-CoV-2 main protease (PDB IDs SRVF and 6Y2E). The profile is shown in magenta lines. The positions of the residues undergoing mutations are shown in red filled circles (A7, G15, M17, V20, T45, D48, M49, R60, K61, A70, G71, L89, K90, P99, Y101, R105, P108, A116, A129, P132, T135, I136, N151, V157, C160, A173, P184, T190, A191, A193, T196, T198, T201, L220, L232, A234, K236, Y237, D248, A255, T259, A260, V261, A266, N274, R279, and S301). (B) PCA analysis of functional dynamics SARS-CoV-2 main protease structures. The slow-mode shapes are shown as mean square fluctuations averaged over the first five lowest frequency modes (in magenta lines). The residues undergoing mutations are shown in red filled circles as in panel A. (C) Structural map of the functional dynamics profiles derived from CG-MD simulations in the SARS-CoV-2 main protease (PDB IDs SRVF and 6Y2E). The color gradient from blue to red indicates the decreasing structural rigidity and increasing flexibility as averaged over the first five low frequency modes. The positions of the residues undergoing mutations are shown in spheres colored according to their level of rigidity/flexibility in slow modes (blue-rigid, red-flexible). The locations of the protease domains I, II, and III are indicated. The catalytic residues HIS41 and CYS145 are shown in sticks. (D) Rotated view for the structural map of functional dynamics profiles in the SARS-CoV-2 main protease with sites of mutations in spheres.

a single unit most likely to maintain the integrity of the catalytic surface. From the MD simulations, we find that domain III is generally not positively correlated with any of the other two domains, can even be negatively correlated with itself on the alternate chain, and may suggest a degree of independence for domain III in terms of dynamics and possibly function as well, at least in its dimeric apo form.

3.10. Coarse-Grained Simulations of SARS-CoV-2 M^{Pro} Structures Reveal Relationships between Mutational Patterns and Functional Motions. To complement all-atom MD simulations and obtain a more granular description of structural dynamics in studied systems, we performed coarse-grained (CG) simulations of the SARS-CoV-2 main protease structures in the free and ligand-bound forms using

the CABS approach^{67–71} (Figure 18). By using a large number of independent CG simulations, we obtained conformational dynamics profiles for the studied systems and analyzed these distributions in the context of examined mutations.

RMSF of protein residues revealed the distribution of stable and flexible regions, thereby allowing the assessment of the extent of mobility for mutational sites (Figure 18, panel A). In domain I, the most flexible residues are in the region 45–75, while for domain II, the L2 loop (residues 165–172) and L3 (residues 185–200) located around the substrate binding pocket also harbor a significant number of flexible positions. In addition, we observed significant fluctuations for surface residues 153–155 and 274–277 (Figure 18, panel A). Of particular interest is the distribution of stable and flexible

residues in the substrate binding pocket. Some pocket residues from domain I (T24, T25, T26, and L27) experience moderate fluctuations, while several other sites (M49 and Y54) displayed considerably higher mobility. Another group of the substrate binding site residues from domain II (S139, F140, and Q189) also exhibited appreciable fluctuations, while residues G143, S144, H164, H163, E166, P168, and C145 showed only moderate changes and remained stable in CG simulations (Figure 18, panel A). Notably and as expected, the catalytic residues C145 and H41 from the substrate binding site also remained stable. The analysis generally showed that domain II residues were stable, while domain III (residues 198–303) showed more flexibility, especially in the peripheral solvent-exposed regions. This domain is involved in regulation of the dimerization through a salt bridge interaction between GLU290 of one protomer and ARG4 of the other protomer. Importantly, we found that these residues remained extremely stable in simulations. Interestingly, buried positions subjected to mutations exhibited different levels of flexibility. While positions A7, V20, A116, A129, T135, I136, V157, C160, A173, and T201 were very stable, other buried sites with registered mutations in domain III (A234 and A266) showed larger fluctuations. It is worth mentioning that the interfacial residue A7 in the N-finger region important for enzymatic activity showed an extreme level of rigidity. Simulation-derived residue–residue couplings were evaluated using principal component analysis (PCA). By comparing slow-mode profiles, we found that functionally significant patterns can be yielded with up to the five slowest eigenvectors that account for 90% of the total variance of the dynamic fluctuations. The functional dynamics profile averaged over the five slow modes showed that domain I (residues 10–99) and domain III (residues 198–303) are mostly mobile in functional motions and can undergo large structural changes (Figure 18, panel B). At the same time, domain II (residues 100–182) is mostly stable during functional dynamics. The distribution of mutational sites clearly indicated the existence of two major clusters. One cluster of mutations is located in highly mobile regions of domain III (T198I, T201A, L220F, L232F, A234V, K236R, Y237H, D248E, A255V, T259I, A260V, V261A, A266V, N274D, R279C, and S301L). These residues involved in protein motion are likely under different evolutionary constraints compared to other functional sites. Another cluster of mutations is distributed in domain II and includes 3 subgroups: a group of fully immobilized positions (A116V, A129V, P132L, T135I, and I136V), a group of bridging (hinge-like) sites that connect rigid and flexible regions (Y101C, R105H, P108S, N151D, V157I/L, C160S, A173V, and P184L/S), and a group of mostly mobile residues (T190I, A191V, and A193V) (Figure 18, panel B). The group of potential hinge sites may be important for controlling regulatory motions, and mutations in these regions (such as V157I/L and P184L/S) may affect global movements in the protease and its enzymatic activity. It is particularly important to dissect the connection between the function of some key residues and their contribution in collective movements. The dimerization residues (R4, M6, S10, G11, E14, N28, S139, F140, S147, E166, E290, and R298) are characterized by different local flexibilities but tend to correspond to low moving regions of the protein in collective motions (Figure 18, panels C and D). The key substrate binding residues (H163, H164, M165, E166, and L167) are located at the very border of structurally immobilized and more flexible regions, and as

such may constitute a hinge region that controls cooperative movements. Notably, some other binding site residues D187, R188, Q189, T190, and A191 are more flexible in slow modes and may undergo functional motions. Substrate recognition sites tend to exhibit structural flexibility and sequence variations so as to enable specific recognition required for mediating substrate specificity. We also explored the functional dynamics profile of the ligand-bound protease complex (Figure 19). This structural map clearly illustrated that the ligand

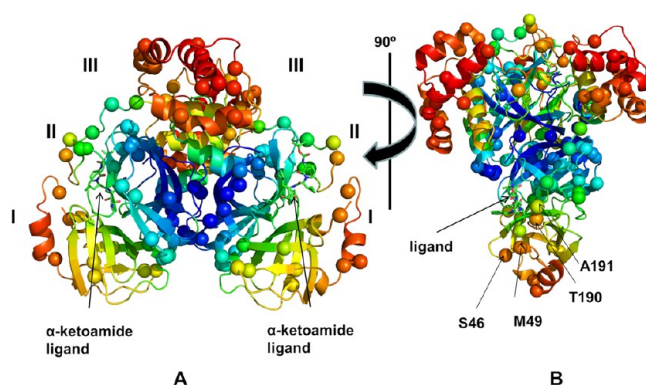


Figure 19. Structural map of functional motion profiles of the SARS-CoV-2 main protease structure complex with a ligand. (A) Structural map of the functional dynamics profiles derived from CG-MD simulations in the SARS-CoV-2 main protease in the complex with the α -ketoamide ligand (PDB ID 6Y2F). The slow-mode shapes are averaged over the first five lowest frequency modes. The color gradient from blue to red indicates the decreasing structural rigidity and increasing flexibility as averaged over the first five low frequency modes. The positions of the residues undergoing mutations (A7, G15, M17, V20, T45, D48, M49, R60, K61, A70, G71, L89, K90, P99, Y101, R105, P108, A116, A129, P132, T135, I136, N151, V157, C160, A173, P184, T190, A191, A193, T196, T198, T201, L220, L232, A234, K236, Y237, D248, A255, T259, A260, V261, A266, N274, R279, and S301) are shown in spheres colored according to their level of rigidity and flexibility in the low frequency modes (blue-rigid, red-flexible). The locations of the protease domains I, II, and III are indicated. The α -ketoamide ligands are shown in sticks in both protomers. (B) Rotated view for the structural map of functional dynamics profiles in the SARS-CoV-2 main protease with sites of mutations in spheres. The position of the α -ketoamide ligand is shown in sticks. The mobile residues in the slow modes from the substrate binding site that form interactions with the ligand (S46, M49, T190, and A191) are indicated by arrows and annotations.

binding site comprised both rigid and flexible residues and was located in the region that bridges an area of high and low structural stability. In particular, we highlighted that residues S46, M49, T190, and A191 in the substrate recognition site and in the ligand proximity may belong to moving regions in the global motions (Figure 19, panel B).

Our analysis shows that structural clusters of mutations may be distinguished by their evolution propensity and global mobility in slow-mode regions. The mobile residues may be predisposed to serve as substrate recognition sites, whereas residues acting as global hinges during collective dynamics are often supported by conserved residues. The observed conservation and mutational patterns may thus be determined by functional catalytic requirements, structural stability and geometrical constraints, and functional dynamics patterns. We found that some sites and corresponding mutations may be associated with dynamic hinge function. The mutability of

hinge sites (Y101C, R105H, P108S, V157I/L, C160S, A173V, and P184L/S) and nearby sites (T190I, A191V, and A193V) may be related with their structural and dynamic signatures to reside in the exposed protein regions rather than in the more conserved protein core. We could also conclude that these sites are located near the active site and control the bending motions needed for catalysis; so, their mutability may have an important functional role for enzymatic activity especially when combined with mutations in adjacent regions.

4. CONCLUSIONS

COVID-19 represents a significant global threat for which no effective solution currently exists. Physical distancing¹¹⁵ has significantly contributed in slowing down the viral progression. However, time is of the essence to find a cure that will counter current and future impacts of the virus, especially for those at a higher risk and for the world economies. Analyzing the structural and dynamic properties of the novel mutants of the SARS-CoV-2 M^{Pro} and contrasts with known SARS-CoV homolog facts gives important pointers and details about its dynamic behavior.

From this work, several missense mutations were found across all domains of the SARS-CoV-2 M^{Pro}, with various levels of support. There are various single residue substitutions, among which several are substitutions of alanine and valine. These mutations have occurred both in buried and solvent-accessible surfaces. From our filtered data set, residue positions 15, 157, and 184 appear to have mutated more than once. A relatively high number of titratable amino acids present in M^{Pro}, which are similar to SARS-CoV, may play an important role in influencing its behavior at various pH levels. Higher backbone flexibility was observed for the isolates EPI_ISL_416720, EPI_ISL_426097, EPI_ISL_421763, EPI_ISL_420610, EPI_ISL_425284, EPI_ISL_421380, EPI_ISL_423772, and EPI_ISL_425886 mainly due to domain III motions and to some extent part of domain I. More generally, regions of lowest flexibility (and high BC) were core residues, while solvent-exposed loops were most flexible. All samples displayed a slight interprotomer twisting motion and domain rotational motions. A high degree of variation was observed in (1) the angle formed between domains I, II, and III, (2) the substrate binding pocket R_g, (3) the interprotomer pocket R_g, and (4) the N-finger flexibility, which may all be good descriptors for characterizing M^{Pro} dynamics. BC values were very similar across all samples, with extreme values being essentially anti-correlated to RMSF. Residues 17 and 128 appear to be very central residues based on the BC network metric, and it is likely that mutations altering their physicochemical property may have the potential to alter dimer stability. The D48E variant (from sample EPI_ISL_425242) leads to a novel "TSEEMLN" motif at the substrate binding flap, which may have repercussions on the efficiency and specificity of substrate binding. Inactive apo M^{Pro} did not show signs of dissociation. A non-canonical pose for the PHE140 widens the accessible surface of the substrate binding pocket by pulling on the oxyanion loop. We propose the presence of a mirrored allosteric interprotomer pocket, supported by two cavity detection approaches, and correlations in compaction dynamics between the interprotomer pocket and the substrate binding pocket. The mirrored pocket may have the potential to accommodate compounds of low molecular weight and polarity. Asymmetries to partial symmetries in R_g distributions were seen for each of the

substrate binding pockets and the interprotomer cavities for each isolate. The compactions of the allosteric subpockets do not seem to noticeably affect the movement of domain III. However, a large portion of the samples displayed overall positive correlations according to DCC, indicating that the mutants generally behave similarly. In each individual DCC plot, domains I and II are found to behave as a single unit, while domain III is generally more independent. Thorough, independent CG simulations of the apo and the ligand-bound M^{Pro} further revealed a connection between regions accumulating clusters of mutations and their degree of residue fluctuation from the slowest modes. Additionally, we report of a possible set of dynamic hinging residues and their tendency to acquire mutations in exposed protein regions while being grounded by less mutable core residues.

As a final note, it is important to be aware that there is an inherent lack of sampling depth in this current analysis of COVID-19 sequences due to the existence of undiagnosed mutations that may be present among infected individuals¹¹⁶ at the time of writing, and in our case, we might have missed certain mutations using our filtering criteria, in our effort to balance accuracy and the number of high confidence representative samples. Therefore, frequencies should be handled with caution as the sampling was done at an early stage of the pandemic.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at <https://pubs.acs.org/doi/10.1021/acs.jcim.0c00634>.

Heat maps of RMSF (Figure S1) and average BC (Figure S2), acknowledgement table detailing sample information (Table S1), and another table (Table S2) showing the high BC residues for each chain (PDF)

■ AUTHOR INFORMATION

Corresponding Author

Özlem Tastan Bishop – Research Unit in Bioinformatics, Department of Microbiology and Biochemistry, Rhodes University, Grahamstown 6140, South Africa; orcid.org/0000-0001-6861-7849; Email: o.tastanbishop@ru.ac.za

Authors

Olivier Sheik Amamuddy – Research Unit in Bioinformatics, Department of Microbiology and Biochemistry, Rhodes University, Grahamstown 6140, South Africa

Gennady M. Verkhivker – Graduate Program in Computational and Data Sciences, Schmid College of Science and Technology, Chapman University, Orange, California 92866, United States; Department of Biomedical and Pharmaceutical Sciences, Chapman University School of Pharmacy, Irvine, California 92618, United States; Department of Pharmacology, Skaggs School of Pharmacy and Pharmaceutical Sciences, University of California San Diego, La Jolla, California 92093, United States; orcid.org/0000-0002-4507-4471

Complete contact information is available at: <https://pubs.acs.org/doi/10.1021/acs.jcim.0c00634>

Funding

O.S.A. is funded as a postdoctoral fellow by H3ABioNet, which is supported by the National Institutes of Health Common

Fund grant number U24HG006941. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. The content of this publication is solely the responsibility of the authors and does not necessarily represent the official views of the funders.

Notes

The authors declare no competing financial interest. Genetic sequences are not disclosed in this article due to the restrictions from GISAID but are fully accessible from the database.

ACKNOWLEDGMENTS

We would like to thank the Center for High Performance Computing for providing computer time for the simulations. We also thank GISAID for making available the COVID-19 genomic material, together with the originating and submitting laboratories (acknowledged in the Supporting Information Table S1).

REFERENCES

- (1) Zhou, P.; Yang, X.-L.; Wang, X.-G.; Hu, B.; Zhang, L.; Zhang, W.; Si, H.-R.; Zhu, Y.; Li, B.; Huang, C.-L.; Chen, H.-D.; Chen, J.; Luo, Y.; Guo, H.; Jiang, R.-D.; Liu, M. Q.; Chen, Y.; Shen, X.-R.; Wang, X.; Zheng, X.-S.; Zhao, K.; Chen, Q.-J.; Deng, F.; Liu, L.-L.; Yan, B.; Zhan, F.-X.; Wang, Y.-Y.; Xiao, G.-F.; Shi, Z.-L. A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature* **2020**, *579*, 270–273.
- (2) Dai, W.; Zhang, B.; Jiang, X.-M.; Su, H.; Li, J.; Zhao, Y.; Xie, X.; Jin, Z.; Liu, F.; Li, C.; Li, Y.; Bai, F.; Wang, H.; Cheng, X.; Cen, X.; Hu, S.; Yang, X.; Wang, J.; Liu, X.; Xiao, G.; Jiang, H.; Rao, Z.; Zhang, L.-K.; Xu, Y.; Yang, H.; Liu, H. Structure-based design of antiviral drug candidates targeting the SARS-CoV-2 main protease. *Science* **2020**, 1331.
- (3) Korber, B.; Fischer, W. M.; Gnanakaran, S.; Yoon, H.; Theiler, J.; Abfalterer, W.; Hengartner, N.; Giorgi, E. E.; Bhattacharya, T.; Foley, B.; Hastie, K. M.; Parker, D. G.; Partridge, D. G.; Evans, C. M.; Freeman, T. M.; de Silva, T. I. Tracking changes in SARS-CoV-2 Spike: evidence that D614G increases infectivity of the COVID-19 virus. *Cell* **2020**, 812.
- (4) Varga, Z.; Flammer, A. J.; Steiger, P.; Haberecker, M.; Andermatt, R.; Zinkernagel, A. S.; Mehra, M. R.; Schuepbach, R. A.; Ruschitzka, F.; Moch, H. Endothelial cell infection and endotheliitis in COVID-19. *The Lancet* **2020**, *395*, 1417–1418.
- (5) Richardson, S.; Hirsch, J. S.; Narasimhan, M.; Crawford, J. M.; McGinn, T.; Davidson, K. W.; Barnaby, D. P.; Becker, L. B.; Chelico, J. D.; Cohen, S. L.; Cookingham, J.; Coppa, K.; Diefenbach, M. A.; Dominello, A. J.; Duer-Hefele, J.; Falzon, L.; Gitlin, J.; Hajizadeh, N.; Harvin, T. G.; Hirschwerk, D. A.; Kim, E. J.; Kozel, Z. M.; Marrast, L. M.; Mogavero, J. N.; Osorio, G. A.; Qiu, M.; Zanos, T. P. Presenting Characteristics, Comorbidities, and Outcomes Among 5700 Patients Hospitalized With COVID-19 in the New York City Area. *JAMA* **2020**, *323*, 2052–2059.
- (6) Helmy, Y. A.; Fawzy, M.; Elswad, A.; Sobieh, A.; Kenney, S. P.; Shehata, A. A. The COVID-19 Pandemic: A Comprehensive Review of Taxonomy, Genetics, Epidemiology, Diagnosis, Treatment, and Control. *Journal of Clinical Medicine* **2020**, *9*, 1225–1225.
- (7) Horby, P.; Lim, W. S.; Emberson, J.; Mafham, M.; Bell, J.; Linsell, L.; Staplin, N.; Brightling, C.; Ustianowski, A.; Elmahi, E.; Prudon, B.; Green, C.; Felton, T.; Chadwick, D.; Rege, K.; Fegan, C.; Chappell, L. C.; Faust, S. N.; Jaki, T.; Jeffery, K.; Montgomery, A.; Rowan, K.; Juszczak, E.; Baillie, J. K.; Haynes, R.; Landray, M. J.; Group, R. C. Dexamethasone in Hospitalized Patients with COVID-19 — Preliminary Report. *N. Engl. J. Med.* **2020**, NEJMoa2021436.
- (8) Beigel, J. H.; Tomashek, K. M.; Dodd, L. E.; Mehta, A. K.; Zingman, B. S.; Kalil, A. C.; Hohmann, E.; Chu, H. Y.; Luetkemeyer, A.; Kline, S.; Lopez de Castilla, D.; Finberg, R. W.; Dierberg, K.; Tanson, V.; Hsieh, L.; Patterson, T. F.; Paredes, R.; Sweeney, D. A.; Short, W. R.; Touloumi, G.; Lye, D. C.; Ohmagari, N.; Oh, M.-d.; Ruiz-Palacios, G. M.; Benfield, T.; Fätkenheuer, G.; Kortepeter, M. G.; Atmar, R. L.; Creech, C. B.; Lundgren, J.; Babiker, A. G.; Pett, S.; Neaton, J. D.; Burgess, T. H.; Bonnett, T.; Green, M.; Makowski, M.; Osinusi, A.; Nayak, S.; Lane, H. C. Remdesivir for the Treatment of Covid-19 — Preliminary Report. *N. Engl. J. Med.* **2020**, DOI: 10.1056/NEJMoa2007764.
- (9) World Health Organization Draft landscape of COVID-19 candidate vaccines; 7 July 2020. 2020; <https://www.who.int/who-documents-detail/draft-landscape-of-covid-19-candidate-vaccines>.
- (10) Rothan, H. A.; Byrareddy, S. N. The epidemiology and pathogenesis of coronavirus disease (COVID-19) outbreak. *J. Autoimmun.* **2020**, *109*, 102433.
- (11) Van Lancker, W.; Parolin, Z. COVID-19, school closures, and child poverty: a social crisis in the making. *Lancet Public Health* **2020**, *5*, e243–e244.
- (12) Elbe, S.; Buckland-Merrett, G. Data, disease and diplomacy: GISAID's innovative contribution to global health. *Global Challenges* **2017**, *1*, 33–46.
- (13) Hatcher, E. L.; Zhdanov, S. A.; Bao, Y.; Blinkova, O.; Nawrocki, E. P.; Ostapchuck, Y.; Schäffer, A. A.; Brister, J. R. Virus Variation Resource — improved response to emergent viral outbreaks. *Nucleic Acids Res.* **2017**, *45*, D482–D490.
- (14) Duffy, S. Why are RNA virus mutation rates so damn high? *PLOS Biol.* **2018**, *16*, e3000003.
- (15) Pond, S. L. K.; Smith, D. M. Are All Subtypes Created Equal? The Effectiveness of Antiretroviral Therapy against Non-Subtype B HIV-1. *Clin. Infect. Dis.* **2009**, *48*, 1306–1309.
- (16) Abecasis, A. B.; Wensing, A. M. J.; Paraskevis, D.; Vercauteren, J.; Theys, K.; Van de Vijver, D. A. M. C.; Albert, J.; Asjö, B.; Balotta, C.; Beshkov, D.; Camacho, R. J.; Clotet, B.; De Gascun, C.; Griskevicius, A.; Grossman, Z.; Hamouda, O.; Horban, A.; Kolupajeva, T.; Korn, K.; Kostrikis, L. G.; Kücherer, C.; Liitsola, K.; Linka, M.; Nielsen, C.; Otelea, D.; Paredes, R.; Poljak, M.; Puchhammer-Stöckl, E.; Schmit, J.-C.; Sönnnerborg, A.; Stanekova, D.; Stanojevic, M.; Struck, D.; Boucher, C. A. B.; Vandamme, A.-M. HIV-1 subtype distribution and its demographic determinants in newly diagnosed patients in Europe suggest highly compartmentalized epidemics. *Retrovirology* **2013**, *10*, 7.
- (17) Sevajol, M.; Subissi, L.; Decroly, E.; Canard, B.; Imbert, I. Insights into RNA synthesis, capping, and proofreading mechanisms of SARS-coronavirus. *Virus Res.* **2014**, *194*, 90–99.
- (18) Pachetti, M.; Marini, B.; Benedetti, F.; Giudici, F.; Mauro, E.; Storici, P.; Masciovecchio, C.; Angeletti, S.; Ciccozzi, M.; Gallo, R. C.; Zella, D.; Ippodrino, R. Emerging SARS-CoV-2 mutation hot spots include a novel RNA-dependent-RNA polymerase variant. *J. Transl. Med.* **2020**, *18*, 179.
- (19) Hadfield, J.; Megill, C.; Bell, S. M.; Huddleston, J.; Potter, B.; Callender, C.; Sagulenko, P.; Bedford, T.; Neher, R. A. Nextstrain: real-time tracking of pathogen evolution. *Bioinformatics* **2018**, *34*, 4121–4123.
- (20) Liang, Y.; Wang, M.-L.; Chien, C.-S.; Yarmishyn, A. A.; Yang, Y.-P.; Lai, W.-Y.; Luo, Y.-H.; Lin, Y.-T.; Chen, Y.-J.; Chang, P.-C.; Chiou, S.-H. Highlight of Immune Pathogenic Response and Hematopathologic Effect in SARS-CoV, MERS-CoV, and SARS-CoV-2 Infection. *Front. Immunol.* **2020**, *11*, 1022.
- (21) Mousavizadeh, L.; Ghasemi, S. Genotype and phenotype of COVID-19: Their roles in pathogenesis. *J. Microbiol., Immunol. Infect.*; **2020**, DOI: 10.1016/j.jmii.2020.03.022.
- (22) Aanouz, I.; Belhassan, A.; El-Khatibi, K.; Lakhli, T.; El-Idrissi, M.; Bouachrine, M. Moroccan Medicinal plants as inhibitors against SARS-CoV-2 main protease: Computational investigations. *J. Biomol. Struct. Dyn.* **2020**, *0*, 1–9.
- (23) Hu, X.; Cai, X.; Song, X.; Li, C.; Zhao, J.; Luo, W.; Zhang, Q.; Ekumi, I. O.; He, Z. Possible SARS-coronavirus 2 inhibitor revealed by simulated molecular docking to viral main protease and host toll-like receptor. *Future Virol.* **2020**, *15* DOI: 10.2217/fvl-2020-0099.
- (24) Peele, K. A.; Potla Durthi, C.; Srihansa, T.; Krupanidhi, S.; Ayyagari, V. S.; Babu, D. J.; Indira, M.; Reddy, A. R.; Venkateswarulu,

T. Molecular docking and dynamic simulations for antiviral compounds against SARS-CoV-2: A computational study. *Inf. Med. Unlocked* **2020**, *19*, 100,345.

(25) Yu, R.; Chen, L.; Lan, R.; Shen, R.; Li, P. Computational screening of antagonists against the SARS-CoV-2 (COVID-19) coronavirus by molecular docking. *Int. J. Antimicrob. Agents* **2020**, *56*, 106012.

(26) Gentile, D.; Patamia, V.; Scala, A.; Sciortino, M. T.; Piperno, A.; Rescifina, A. Putative Inhibitors of SARS-CoV-2 Main Protease from A Library of Marine Natural Products: A Virtual Screening and Molecular Modeling Study. *Marine Drugs* **2020**, *18*, 225.

(27) Wu, C.; Liu, Y.; Yang, Y.; Zhang, P.; Zhong, W.; Wang, Y.; Wang, Q.; Xu, Y.; Li, M.; Li, X.; Zheng, M.; Chen, L.; Li, H. Analysis of therapeutic targets for SARS-CoV-2 and discovery of potential drugs by computational methods. *Acta Pharm. Sin. B* **2020**, *10*, 766–788.

(28) Alazmi, M.; Motwalli, O. In silico virtual screening, characterization, docking and molecular dynamics studies of crucial SARS-CoV-2 proteins. *J. Biomol. Struct. Dyn.* **2020**, *0*, 1–11.

(29) Kumar, S.; Sharma, P. P.; Shankar, U.; Kumar, D.; Joshi, S. K.; Pena, L.; Durvasula, R.; Kumar, A.; Kempaiah, P.; Poonam; Rathi, B. Discovery of New Hydroxyethylamine Analogs against 3CL^{pro} Protein Target of SARS-CoV-2: Molecular Docking, Molecular Dynamics Simulation, and Structure–Activity Relationship Studies. *J. Chem. Inf. Model.* **2020**, DOI: 10.1021/acs.jcim.0c00326.

(30) Ghosh, R.; Chakraborty, A.; Biswas, A.; Chowdhuri, S. 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.* **2020**, 1–14.

(31) Khan, M. T.; Ali, A.; Wang, Q.; Irfan, M.; Khan, A.; Zeb, M. T.; Zhang, Y.-J.; Chinnasamy, S.; Wei, D.-Q. Marine natural compounds as potents inhibitors against the main protease of SARS-CoV-2—a molecular dynamic study. *J. Biomol. Struct. Dyn.* **2020**, *0*, 1–11.

(32) Khelifaoui, H.; Harkati, D.; Saleh, B. A. Molecular docking, molecular dynamics simulations and reactivity, studies on approved drugs library targeting ACE2 and SARS CoV-2 binding with ACE2. *J. Biomol. Struct. Dyn.* **2020**, *0*, 1–17.

(33) Alamri, M. A.; Tahir ul Qamar, M.; Mirza, M. U.; Bhadane, R.; Alqahtani, S. M.; Muneer, I.; Froeyen, M.; Salo-Ahen, O. M. H. Pharmacoinformatics and molecular dynamics simulation studies reveal potential covalent and FDA-approved inhibitors of SARS-CoV-2 main protease 3CL^{pro}. *J. Biomol. Struct. Dyn.* **2020**, *0*, 1–13.

(34) Tahir ul Qamar, M.; Alqahtani, S. M.; Alamri, M. A.; Chen, L.-L. Structural basis of SARS-CoV-2 3CL^{pro} and anti-COVID-19 drug discovery from medicinal plants. *J. Pharm. Anal.* **2020**, 313–319.

(35) Mahmud, S.; Uddin, M. A. R.; Zaman, M.; Sujon, K. M.; Rahman, M. E.; Shehab, M. N.; Islam, A.; Alom, M. W.; Amin, A.; Akash, A. S.; Saleh, M. A. Molecular docking and dynamics study of natural compound for potential inhibition of mainprotease of SARS-CoV-2. *J. Biomol. Struct. Dyn.* **2020**, *0*, 1–9.

(36) Suárez, D.; Díaz, N. SARS-CoV-2 Main Protease: A Molecular Dynamics Study. *J. Chem. Inf. Model.* **2020**, DOI: 10.1021/acs.jcim.0c00575.

(37) Cava, C.; Bertoli, G.; Castiglioni, I. In silico discovery of candidate drugs against covid-19. *Viruses* **2020**, *12*, 404.

(38) Gysi, D. M.; Valle, I. D.; Zitnik, M.; Ameli, A.; Gan, X.; Varol, O.; Sanchez, H.; Baron, R. M.; Ghiassian, S. D.; Patten, J. J.; Davey, R.; Loscalzo, J.; Barabási, A.-L. Network Medicine Frame work for Identifying Drug Repurposing Opportunities for COVID-19. *arXiv preprint arXiv: 2020, 2004, 07229*.

(39) Joshi, R. S.; Jagdale, S. S.; Bansode, S. B.; Shankar, S. S.; Tellis, M. B.; Pandya, V. K.; Chugh, A.; Giri, A. P.; Kulkarni, M. J. Discovery of potential multi-target-directed ligands by targeting host-specific SARS-CoV-2 structurally conserved main protease. *J. Biomol. Struct. Dyn.* **2020**, 1–16.

(40) Zhou, Y.; Hou, Y.; Shen, J.; Huang, Y.; Martin, W.; Cheng, F. Network-based drug repurposing for novel coronavirus 2019-nCoV/SARS-CoV-2. *Cell Discovery* **2020**, *6*, 14.

(41) Das, S.; Sarmah, S.; Lyndem, S.; Singha Roy, A. An investigation into the identification of potential inhibitors of SARS-CoV-2 main protease using molecular docking study. *J. Biomol. Struct. Dyn.* **2020**, 1–11.

(42) Jin, Z.; Du, X.; Xu, Y.; Deng, Y.; Liu, M.; Zhao, Y.; Zhang, B.; Li, X.; Zhang, L.; Peng, C.; Duan, Y.; Yu, J.; Wang, L.; Yang, K.; Liu, F.; Jiang, R.; Yang, X.; You, T.; Liu, X.; Yang, X.; Bai, F.; Liu, H.; Liu, X.; Guddat, L. W.; Xu, W.; Xiao, G.; Qin, C.; Shi, Z.; Jiang, H.; Rao, Z.; Yang, H. Structure of Mpro from COVID-19 virus and discovery of its inhibitors. *Nature* **2020**, 289.

(43) Shah, B.; Modi, P.; Sagar, S. R. In silico studies on therapeutic agents for COVID-19: Drug repurposing approach. *Life Sci.* **2020**, *252*, 117,652.

(44) Randhawa, G. S.; Soltysiak, M. P. M.; El Roz, H.; de Souza, C. P. E.; Hill, K. A.; Kari, L. Machine learning using intrinsic genomic signatures for rapid classification of novel pathogens: COVID-19 case study. *PLoS ONE* **2020**, *15*, e0232391.

(45) Ton, A.-T.; Gentile, F.; Hsing, M.; Ban, F.; Cherkasov, A. Rapid Identification of Potential Inhibitors of SARS-CoV-2 Main Protease by Deep Docking of 1.3 Billion Compounds. *Mol. Inf.* **2020**, 1–7.

(46) Anand, K.; Ziebuhr, J.; Wadhani, P.; Mesters, J. R.; Hilgenfeld, R. Coronavirus Main Proteinase (3CL^{pro}) Structure: Basis for Design of Anti-SARS Drugs. *Science* **2003**, *300*, 1763–1767.

(47) Chen, Y. W.; Yiu, C.-P. B.; Wong, K.-Y. Prediction of the SARS-CoV-2 (2019-nCoV) 3C-like protease (3CL^{pro}) structure: virtual screening reveals velpatasvir, ledipasvir, and other drug repurposing candidates. *F1000Research* **2020**, *9*, 129.

(48) Yang, H.; Xie, W.; Xue, X.; Yang, K.; Ma, J.; Liang, W.; Zhao, Q.; Zhou, Z.; Pei, D.; Ziebuhr, J.; Hilgenfeld, R.; Yuen, K. Y.; Wong, L.; Gao, G.; Chen, S.; Chen, Z.; Ma, D.; Bartlam, M.; Rao, Z. Design of Wide-Spectrum Inhibitors Targeting Coronavirus Main Proteases. *PLoS Biol.* **2005**, *3*, No. e324.

(49) Xue, X.; Yu, H.; Yang, H.; Xue, F.; Wu, Z.; Shen, W.; Li, J.; Zhou, Z.; Ding, Y.; Zhao, Q.; Zhang, X. C.; Liao, M.; Bartlam, M.; Rao, Z. Structures of Two Coronavirus Main Proteinases: Implications for Substrate Binding and Antiviral Drug Design. *J. Virol.* **2008**, *82*, 2515–2527.

(50) 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**, *412*, 409.

(51) Chen, S.; Hu, T.; Zhang, J.; Chen, J.; Chen, K.; Ding, J.; Jiang, H.; Shen, X. Mutation of Gly-11 on the Dimer Interface Results in the Complete Crystallographic Dimer Dissociation of Severe Acute Respiratory Syndrome Coronavirus 3C-like Protease. *J. Biol. Chem.* **2008**, *283*, 554–564.

(52) Chen, H.; Wei, P.; Huang, C.; Tan, L.; Liu, Y.; Lai, L. Only One Protomer Is Active in the Dimer of SARS 3C-like Proteinase. *J. Biol. Chem.* **2006**, *281*, 13894–13898.

(53) Anand, K.; Palm, G. J.; Mesters, J. R.; Siddel, S. G.; Ziebuhr, J.; Hilgenfeld, R. Structure of coronavirus main proteinase reveals combination of a chymotrypsin fold with an extra α -helical domain. *EMBO J.* **2002**, *21*, 3213–3224.

(54) 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. *J. Biol. Chem.* **2004**, *279*, 24765–24773.

(55) Shi, J.; Song, J. The catalysis of the SARS 3C-like protease is under extensive regulation by its extra domain. *FEBS J.* **2006**, *273*, 1035–1045.

(56) Brown, D. K.; Penkler, D. L.; Amamuddy, O. S.; Ross, C.; Atilgan, A. R.; Atilgan, C.; Bishop, Ö. T. MD-TASK: a software suite for analyzing molecular dynamics trajectories. *Bioinformatics* **2017**, *33*, 2768–2771.

(57) Fearon, D.; Owen, C.; Douangamath, A.; Lukacik, P.; Powell, A.; Strain-Damerell, C.; Resnick, E.; Krojer, T.; Gehrtz, P.; Wild, C.; Aimon, A.; Brandao-Neto, J.; Carbery, A.; Dunnett, L.; Skyner, R.; Snee, M.; London, N.; Walsh, M.; von Delft, F. *PanDDA analysis*

group deposition – Crystal Structure of SARS-CoV-2 main protease in complex with PCM-0102306; 2020.

(58) Burley, S. K.; Berman, H. M.; Bhikadiya, C.; Bi, C.; Chen, L.; Di Costanzo, L.; Christie, C.; Dalenberg, K.; Duarte, J. M.; Dutta, S.; Feng, Z.; Ghosh, S.; Goodsell, D. S.; Green, R. K.; Guranović, V.; Guzenko, D.; Hudson, B. P.; Kalro, T.; Liang, Y.; Lowe, R.; Namkoong, H.; Peisach, E.; Periskova, I.; Prlic, A.; Randle, C.; Rose, A.; Rose, P.; Sala, R.; Sekharan, M.; Shao, C.; Tan, L.; Tao, Y.-P.; Valasatava, Y.; Voigt, M.; Westbrook, J.; Woo, J.; Yang, H.; Young, J.; Zhuravleva, M.; Zardecki, C. RCSB Protein Data Bank: biological macromolecular structures enabling research and education in fundamental biology, biomedicine, biotechnology and energy. *Nucleic Acids Res.* **2019**, *47*, D464–D474.

(59) Schrodinger, L. *The PyMOL Molecular Graphics System*; Version 2.4.0a0. 2015.

(60) Camacho, C.; Coulouris, G.; Avagyan, V.; Ma, N.; Papadopoulos, J.; Bealer, K.; Madden, T. L. BLAST+: architecture and applications. *BMC Bioinf.* **2009**, *10*, 421.

(61) Cock, P. J. A.; Antao, T.; Chang, J. T.; Chapman, B. A.; Cox, C. J.; Dalke, A.; Friedberg, I.; Hamelryck, T.; Kauff, F.; Wilczynski, B.; de Hoon, M. J. L. Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* **2009**, *25*, 1422–1423.

(62) Sali, A.; Blundell, T. L. Comparative Protein Modelling by Satisfaction of Spatial Restraints. *J. Mol. Biol.* **1993**, *234*, 779–815.

(63) Dolinsky, T. J.; Nielsen, J. E.; McCammon, J. A.; Baker, N. A. PDB2PQR: An automated pipeline for the setup of Poisson-Boltzmann electrostatics calculations. *Nucleic Acids Res.* **2004**, *32*, W665.

(64) Jubb, H. C.; Higuero, A. P.; Ochoa-Montaño, B.; Pitt, W. R.; Ascher, D. B.; Blundell, T. L. Arpeggio: A Web Server for Calculating and Visualising Interatomic Interactions in Protein Structures. *J. Mol. Biol.* **2017**, *429*, 365–371.

(65) Dassault Systemes BIOVIA *Discovery Studio Visualizer*; v19.1.0.18287. 2019; <http://accelrys.com/products/discovery-studio/>.

(66) Van Der Spoel, D.; Lindahl, E.; Hess, B.; Groenhof, G.; Mark, A. E.; Berendsen, H. J. C. GROMACS: Fast, flexible, and free. *J. Comput. Chem.* **2005**, *26*, 1701–1718.

(67) Kolinski, A. Protein modeling and structure prediction with a reduced representation. *Acta Biochim. Pol.* **2004**, *51*, 349–371.

(68) Kmiecik, S.; Gront, D.; Kolinski, M.; Wieteska, L.; Dawid, A. E.; Kolinski, A. Coarse Grained Protein Models and Their Applications. *Chemical Reviews* **2016**, *116*, 7898–7936.

(69) Kmiecik, S.; Kouza, M.; Badaczewska-Dawid, A.; Kloczkowski, A.; Kolinski, A. Modeling of Protein Structural Flexibility and Large-Scale Dynamics: Coarse-Grained Simulations and Elastic Network Models. *Int. J. Mol. Sci.* **2018**, *19*, 3496.

(70) Ciemny, M.; Badaczewska-Dawid, A. E.; Pikuzinska, M.; Kolinski, A.; Kmiecik, S. Modeling of Disordered Protein Structures Using Monte Carlo Simulations and Knowledge Based Statistical Force Fields. *Int. J. Biences* **2019**, *20*, 606.

(71) Kurcinski, M.; Oleniecki, T.; Ciemny, M. P.; Kuriata, A.; Kolinski, A.; Kmiecik, S. CABS-flex standalone: a simulation environment for fast modeling of protein flexibility. *Bioinformatics* **2019**, *35*, 694–695.

(72) Bakan, A.; Meireles, L. M.; Bahar, I. ProDy: Protein Dynamics Inferred from Theory and Experiments. *Bioinformatics* **2011**, *27*, 1575–1577.

(73) Kozakov, D.; Grove, L. E.; Hall, D. R.; Bohnuud, T.; Mottarella, S. E.; Luo, L.; Xia, B.; Beglov, D.; Vajda, S. The FTMap family of web servers for determining and characterizing ligand-binding hot spots of proteins. *Nature Protocols* **2015**, *10*, 733–755.

(74) Smith, R. H. B.; Dar, A. C.; Schlessinger, A. PyVOL: a PyMOL plugin for visualization, comparison, and volume calculation of drug-binding sites. *bioRxiv* **2019**, 816702.

(75) Hunter, J. D. Matplotlib: A 2D Graphics Environment. *Comput. Sci. Eng.* **2007**, *9*, 90–95.

(76) McKinney, W. Data Structures for Statistical Computing in Python. *Proceedings of the 9th Python in Science Conference*; 2010, pp. 51–56.

(77) Van Der Walt, S.; Colbert, S. C.; Varoquaux, G. The NumPy array: A structure for efficient numerical computation. *Comput. Sci. Eng.* **2011**, *13*, 22–30.

(78) Virtanen, P.; Gommers, R.; Oliphant, T. E.; Haberland, M.; Reddy, T.; Cournapeau, D.; Burovski, E.; Peterson, P.; Weckesser, W.; Bright, J.; van der Walt, S. J.; Brett, M.; Wilson, J.; Millman, K. J.; Mayorov, N.; Nelson, A. R. J.; Jones, E.; Kern, R.; Larson, E.; Carey, C. J.; Polat, I.; Feng, Y.; Moore, E. W.; VanderPlas, J.; Laxalde, D.; Perktold, J.; Cimrman, R.; Henriksen, I.; Quintero, E. A.; Harris, C. R.; Archibald, A. M.; Ribeiro, A. H.; Pedregosa, F.; van Mulbregt, P.; SciPy 1.0 Contributors. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat. Methods* **2020**, *17*, 261–272.

(79) McGibbon, R. T.; Beauchamp, R. A.; Harrigan, M. P.; Klein, C.; Swails, J. M.; Hernández, C. X.; Schwantes, C. R.; Wang, L.-P.; Lane, T. J.; Pande, V. S. MD Traj: A Modern Open Library for the Analysis of Molecular Dynamics Trajectories. *Biophys. J.* **2015**, *109*, 1528–1532.

(80) Nguyen, H.; Case, D. A.; Rose, A. S. NGLview—interactive molecular graphics for Jupyter notebooks. *Bioinformatics* **2018**, *34*, 1241–1242.

(81) Scott, D. W. *Multivariate Density Estimation: Theory, Practice, and Visualization*; John Wiley & Sons: 1992; Chapter 6, pp. 125–193, DOI: 10.1002/9780470316849.ch6.

(82) Rice, A. M.; Morales, A. C.; Ho, A. T.; Mordstein, C.; Mühlhausen, S.; Watson, S.; Cano, L.; Young, B.; Kudla, G.; Hurst, L. D. Evidence for strong mutation bias towards, and selection against, T/U content in SARS-CoV-2: implications for vaccine design. *bioRxiv* **2020**, DOI: 10.1101/2020.05.11.088112.

(83) Rambaut, A.; Holmes, E. C.; Hill, V.; O'Toole, A.; McCrone, J. T.; Ruis, C.; du Plessis, L.; Pybus, O. G. A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. *Nat. Microbiol.* **2020**, 1–14.

(84) Ma, C.; Sacco, M. D.; Hurst, B.; Townsend, J. A.; Hu, Y.; Szeto, T.; Zhang, X.; Tarbet, B.; Marty, M. T.; Chen, Y.; Wang, J. Boceprevir, GC-376, and calpain inhibitors II, XII inhibit SARS-CoV-2 viral replication by targeting the viral main protease. *Cell Res.* **2020**, *30*, 678–692.

(85) Rasschaert, D.; Duarte, M.; Laude, H. Porcine respiratory coronavirus differs from transmissible gastroenteritis virus by a few genomic deletions. *J. Gen. Virol.* **1990**, *71*, 2599–2607.

(86) Carter, R. W.; Sanford, J. C. A new look at an old virus: patterns of mutation accumulation in the human H1N1 influenza virus since 1918. *Theor. Biol. Med. Modelling* **2012**, *9*, 42.

(87) He, J. F.; Peng, G. W.; Min, J.; Yu, D.-W.; Liang, W.-J.; Zhang, S.-Y.; Xu, R.-H.; Zheng, H.-Y.; Wu, X.-W.; Xu, J.; Wang, Z.-H.; Fang, L.; Zhang, X.; Li, H.; Yan, X.-G.; Lu, J.-H.; Hu, Z.-H.; Huang, J.-C.; Wan, Z.-Y.; Hou, J.-L.; Lin, J.-Y.; Song, H.-D.; Wang, S.-Y.; Zhou, X.-J.; Zhang, G.-W.; Gu, B.-W.; Zheng, H.-J.; Zhang, X.-L.; He, M.; Zheng, K.; Wang, B.; Fu, G.; Wang, X.-N.; Chen, S.-J.; Chen, Z.; Hao, P.; Tang, H.; Ren, S.-X.; Zhong, Y.; Guo, Z.-M.; Liu, Q.; Miao, Y.-G.; Kong, X.-Y.; He, W.-Z.; Li, Y.-X.; Wu, C.-I.; Zhao, G.-P.; Chiu, R. W. K.; Chim, S. S. C.; Tong, Y.-K.; Chan, P. K.; Tam, J. S.; Lo, Y. M. D. Molecular Evolution of the SARS Coronavirus, during the Course of the SARS Epidemic in China. *Science* **2004**, *303*, 1666–1669.

(88) Pfeiffer, J. K.; Kirkegaard, K. Increased Fidelity Reduces Poliovirus Fitness and Virulence under Selective Pressure in Mice. *PLoS Pathog.* **2005**, *1*, e11.

(89) Vignuzzi, M.; Stone, J. K.; Andino, R. Ribavirin and lethal mutagenesis of poliovirus: molecular mechanisms, resistance and biological implications. *Virus Research* **2005**, *107*, 173–181.

(90) Wensing, A.; Calvez, V.; Gunthard, H.; Johnson, V.; Paredes, R.; Pillay, D.; Shafer, R.; Richman, D. 2017 Update of the Drug Resistance Mutations in HIV-1. *Top. Antiviral Med.* **2017**, 132–133.

(91) Wensing, A. M.; Calvez, V.; Ceccherini-Silberstein, F.; Charpentier, C.; Gunthard, H.; Paredes, R.; Shafer, S. W.; Richman,

D. D. 2019 Update of the Drug Resistance Mutations in HIV-1. *Top. Antiviral Med.*, 2019, 111.

(92) Brown, D. K.; Amamuddy, O. S.; Bishop, Ö. T. Structure-Based Analysis of Single Nucleotide Variants in the Renin-Angiotensinogen Complex. *Global Heart* 2017, 12, 121.

(93) Jubb, H. C.; Pandurangan, A. P.; Turner, M. A.; Ochoa-Montano, B.; Blundell, T. L.; Ascher, D. B. Mutations at protein-protein interfaces: Small changes over big surfaces have large impacts on human health. *Prog. Biophys. Mol. Biol.* 2017, 128, 3–13.

(94) Sali, A. MODELLER: A Program for Protein Structure Modeling Release. *Rockefeller Univ.* 2013, 9, 779.

(95) Yang, H.; Yang, M.; Ding, Y.; Liu, Y.; Lou, Z.; Zhou, Z.; Sun, L.; Mo, L.; Ye, S.; Pang, H.; Gao, G. F.; Anand, K.; Bartlam, M.; Hilgenfeld, R.; Rao, Z. The crystal structures of severe acute respiratory syndrome virus main protease and its complex with an inhibitor. *Proc. Natl. Acad. Sci. U. S. A.* 2003, 100, 13190–13195.

(96) Humphrey, W.; Dalke, A.; Schulten, K. VMD: Visual molecular dynamics. *J. Mol. Graphics* 1996, 14, 33–38.

(97) Goyal, B.; Goyal, D. Targeting the Dimerization of the Main Protease of Coronaviruses: A Potential Broad-Spectrum Therapeutic Strategy. *ACS Comb. Sci.* 2020, 22, 297–305.

(98) Wei, P.; Fan, K.; Chen, H.; Ma, L.; Huang, C.; Tan, L.; Xi, D.; Li, C.; Liu, Y.; Cao, A.; Lai, L. The N-terminal octapeptide acts as a dimerization inhibitor of SARS coronavirus 3C-like proteinase. *Biochem. Biophys. Res. Commun.* 2006, 339, 865–872.

(99) Chou, C.-Y.; Chang, H.-C.; Hsu, W.-C.; Lin, T.-Z.; Lin, C.-H.; Chang, G.-G. Quaternary Structure of the Severe Acute Respiratory Syndrome (SARS) Coronavirus Main Protease. *Biochemistry* 2004, 43, 14958–14970.

(100) Shi, J.; Sivaraman, J.; Song, J. Mechanism for Controlling the Dimer-Monomer Switch and Coupling Dimerization to Catalysis of the Severe Acute Respiratory Syndrome Coronavirus 3C-Like Protease. *J. Virol.* 2008, 82, 4620–4629.

(101) Chen, S.; Chen, L.; Tan, J.; Chen, J.; Du, L.; Sun, T.; Shen, J.; Chen, K.; Jiang, H.; Shen, X. Severe Acute Respiratory Syndrome Coronavirus 3C-like Proteinase N Terminus Is Indispensable for Proteolytic Activity but Not for Enzyme Dimerization. *J. Biol. Chem.* 2005, 280, 164–173.

(102) Cheng, S.-C.; Chang, G.-G.; Chou, C.-Y. Mutation of Glu-166 Blocks the Substrate Induced Dimerization of SARS Coronavirus Main Protease. *Biophys. J.* 2010, 98, 1327–1336.

(103) Li, C.; Teng, X.; Qi, Y.; Tang, B.; Shi, H.; Ma, X.; Lai, L. Conformational Flexibility of a Short Loop near the Active Site of the SARS-3CLpro is Essential to Maintain Catalytic Activity. *Sci. Rep.* 2016, 6, 20918.

(104) Li, C.; Qi, Y.; Teng, X.; Yang, Z.; Wei, P.; Zhang, C.; Tan, L.; Zhou, L.; Liu, Y.; Lai, L. Maturation mechanism of severe acute respiratory syndrome (SARS) coronavirus 3C like proteinase. *J. Biol. Chem.* 2010, 285, 28134–28140.

(105) Frishman, D.; Argos, P. Knowledge-based protein secondary structure assignment. *Proteins: Struct., Funct., Genet.* 1995, 23, 566–579.

(106) Faraggi, E.; Dunker, A. K.; Jernigan, R. L.; Kloczkowski, A. Entropy, Fluctuations, and Disordered Proteins. *Entropy* 2019, 21, 764.

(107) Horowitz, S.; Trievel, R. C. Carbon–Oxygen Hydrogen Bonding in Biological Structure and Function. *J. Biol. Chem.* 2012, 287, 41576–41582.

(108) Zhao, Q.; Li, S.; Xue, F.; Zou, Y.; Chen, C.; Bartlam, M.; Rao, Z. Structure of the Main Protease from a Global Infectious Human Coronavirus, HCoV-HKU1. *J. Virol.* 2008, 82, 8647–8655.

(109) 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. *Sci. Rep.* 2016, 6, 22677.

(110) Penkler, D. L.; Atilgan, C.; Bishop, Ö. T. Allosteric Modulation of Human Hsp90 α Conformational Dynamics. *J. Chem. Inf. Model.* 2018, 58, 383–404.

(111) Aledo, J. C. Methionine in proteins: The Cinderella of the proteinogenic amino acids. *Protein Science* 2019, 28, 1785–1796.

(112) Hu, T.; Zhang, Y.; Li, L.; Wang, K.; Chen, S.; Chen, J.; Ding, J.; Jiang, H.; Shen, X. Two adjacent mutations on the dimer interface of SARS coronavirus 3C-like protease cause different conformational changes in crystal structure. *Virology* 2009, 388, 324–334.

(113) Zheng, K.; Ma, G.; Zhou, J.; Zen, M.; Zhao, W.; Jiang, Y.; Yu, Q.; Feng, J. Insight into the activity of SARS main protease: Molecular dynamics study of dimeric and monomeric form of enzyme. *Proteins: Struct., Funct., Bioinf.* 2006, 66, 467–479.

(114) Kim, S.; Chen, J.; Cheng, T.; Gindulyte, A.; He, J.; He, S.; Li, Q.; Shoemaker, B. A.; Thiessen, P. A.; Yu, B.; Zaslavsky, L.; Zhang, J.; Bolton, E. E. PubChem 2019 update: improved access to chemical data. *Nucleic Acids Res.* 2019, 47, D1102–D1109.

(115) Chu, D. K.; Akl, E. A.; Duda, S.; Solo, K.; Yaacoub, S.; Schünemann, H. J.; Chu, D. K.; Akl, E. A.; El-harakeh, A.; Bognanni, A.; Lotfi, T.; Loeb, M.; Hajizadeh, A.; Bak, A.; Izcovich, A.; Cuello-Garcia, C. A.; Chen, C.; Harris, D. J.; Borowiack, E.; Chamseddine, F.; Schünemann, F.; Morgano, G. P.; Schünemann, G. E. U. M.; Chen, G.; Zhao, H.; Neumann, I.; Chan, J.; Khabsa, J.; Hneiny, L.; Harrison, L.; Smith, M.; Rizk, N.; Rossi, P. G.; AbiHanna, P.; El-khoury, R.; Stalteri, R.; Baldeh, T.; Piggott, T.; Zhang, Y.; Saad, Z.; Khamis, A.; Reinap, M.; Duda, S.; Solo, K.; Yaacoub, S.; Schünemann, H. J. Physical distancing, face masks, and eye protection to prevent person-to-person transmission of SARS CoV-2 and COVID-19: a systematic review and meta-analysis. *Lancet* 2020, 395, 1973–1987.

(116) Chookajorn, T. Evolving COVID-19 conundrum and its impact. *Proc. Natl. Acad. Sci.* 2020, 117, 12520–12521.