Biochemistry

pubs.acs.org/biochemistry Article

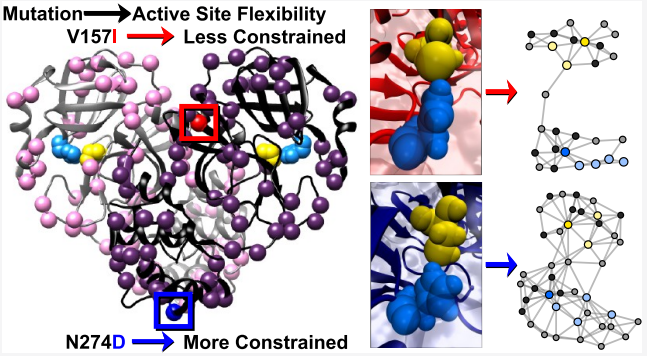
Sequence Characterization and Molecular Modeling of Clinically Relevant Variants of the SARS-CoV-2 Main Protease

Thomas J. Cross, Gemma R. Takahashi, Elizabeth M. Diessner, Marquise G. Crosby, Vesta Farahmand, Shannon Zhuang, Carter T. Butts,* and Rachel W. Martin*

Cite This: https://dx.doi.org/10.1021/acs.biochem.0c00462 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: The SARS-CoV-2 main protease (M^{pro}) is essential to viral replication and cleaves highly specific substrate sequences, making it an obvious target for inhibitor design. However, as for any virus, SARS-CoV-2 is subject to constant neutral drift and selection pressure, with new M^{pro} mutations arising over time. Identification and structural characterization of M^{pro} variants is thus critical for robust inhibitor design. Here we report sequence analysis, structure predictions, and molecular modeling for seventy-nine M^{pro} variants, constituting all clinically observed mutations in this protein as of April 29, 2020. Residue substitution is widely distributed, with some tendency toward larger and more hydrophobic residues. Modeling and protein structure network analysis suggest differences in cohesion and active site flexibility, revealing patterns in viral evolution that have relevance for drug discovery.



Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) emerged in late 2019[1] and rapidly spread worldwide, causing an ongoing pandemic. Although the sequence of its RNA genome is highly similar to that of SARS-CoV-1, SARS-CoV-2 is believed to have arisen independently from a bat coronavirus,[2] to which it shares 96% similarity.[3] The emerging SARS-CoV-2 subsequently gained a modified spike protein due to recombination in an intermediate host, possibly the pangolin,[4,5] followed by purifying selection for binding to the human ACE2 protein.[6] No therapeutic agents able to reduce SARS-CoV-2 mortality in clinical settings are yet known, although extensive efforts are underway to discover new drugs or repurpose existing ones to inhibit key viral proteins. Here we focus on the main protease (M^{pro}), which plays a critical role in viral replication. Like other betacoronaviruses, SARS-CoV-2 is a positive-sense RNA virus that expresses all of its proteins as a single polypeptide chain, which is cleaved by M^{pro} and the papain-like protease PL^{pro} to yield the mature proteins.[7]

Inhibiting this key enzyme would prevent viral replication, reducing viral load and thus symptom intensity. A similar approach was instrumental in making HIV a manageable disease.[8−10] However, the proteins in question differ markedly, rendering HIV protease inhibitors ineffective against SARS-CoV-2; indeed, a standard HIV protease inhibitor combination did not prove effective against COVID-19 in a recent clinical trial.[11] Specifically, HIV protease is an aspartic protease, whereas M^{pro} is a 3CL cysteine protease. The 3CL cysteine proteases are characterized by a chymotrypsin-like fold and a cysteine-histidine catalytic dyad in the active site, implying both different structures and distinct chemical mechanisms. Although the general strategy of seeking protease inhibitors is hence viable for both SARS-CoV-2 and HIV, drug development for the former depends on characterizing this novel enzyme.

The M^{pro} sequence and three-dimensional structure are highly conserved among coronaviruses, with a characteristic three-domain fold.[12] Domains I and II make up the catalytic region, which has a chymotrypsin-like structure, while domain III is an α-helical domain that is primarily responsible for dimerization.[13,14] Although a mix of monomers and dimers is found in solution,[15] studies of the SARS-CoV-1 M^{pro} have shown that the dimer is the functional state in the mature protein, whereas the monomer has dramatically reduced trans-enzymatic activity against the substrate.[13] Removing key residues involved in dimerization by mutation or truncation of the C-terminus results in lower enzymatic activity concomitant with an increased population of monomers.[16] The N-finger (residues 1−7) is not required for dimerization[17] but is necessary for activity, apparently due to highly specific

Received: June 2, 2020
Revised: September 12, 2020
Published: September 15, 2020

interactions between the N-terminus of the inactive monomer and domains II and III of the active one.[18−20] In the G11A variant, the N-finger is unable to occupy the correct binding site, leading to dissociation of the dimer, loss of activity, and collapse of the oxyanion hole that is an essential part of the substrate-binding pocket.[21] The global efforts to document clinically isolated SARS-CoV-2 sequences since the beginning of the pandemic provides an opportunity to examine the impact of mutations on the protein structure and dynamics, and as more sequences are collected going forward, to discover which residues and functional features are absolutely conserved.

Molecular modeling is an important tool for guiding inhibitor discovery, making it possible to evaluate large numbers of candidate drugs *in silico* to select experimental targets; however, standard approaches screen against only one version of the protein, typically the reference or wild-type (WT) sequence. In a host population, mutations accumulate with each viral passage, generating a mutational landscape rather than a single protein. The design of robust inhibitors that can protect against the multiple strains encountered in clinical settings requires the characterization of this sequence space and the populations of conformations it engenders. Furthermore, effective and rapid response to future emerging coronavirus diseases requires both *in silico* screening and experimental testing of antiviral agents and a validated library of relatively general inhibitors that can be used as a basis for the development of specialized therapeutics. Central to the success of that effort will be developing an understanding of structural and functional variation in SARS-CoV-2 proteins, particularly as mutations accumulate and new strains emerge.

In general, SARS-CoV-2 is mutating more slowly than would be expected under neutral drift, suggesting that most of the genome is subject to purifying selection,[22] although this does not rule out adaptive mutations in specific genes or other sequence regions.[23] Especially with a newly emerged zoonotic disease, there is no reason to assume that any viral protein is currently at a global optimum with respect to function, and function-enhancing variants may appear at any time. For example, mutation of the SARS-CoV-2 spike protein is a topic of intense current interest, as its D614G variant appears to confer enhanced infectivity,[24] has spread rapidly through the global population,[25] and correlates with a higher mortality rate.[26] M^{pro} appears to be relatively tolerant of mutations near the active site,[27] underscoring the importance of mapping the mutational landscape of active variants so that inhibitors whose binding depends on highly specific interactions with mutation-prone residues can be eliminated early in the screening process. Here we characterize all 79 known variants of M^{pro} as of 29 April, 2020, and analyze trends in amino acid substitutions and the resulting structural changes for both monomers and dimers, using molecular modeling and moiety-level protein structure network analysis. Our analysis shows a trend toward substitution for larger and more hydrophobic residues versus the WT protein. Analysis of active site networks (ASNs) from M^{pro} variants suggests differences in active site flexibility and cohesion that may serve to guide the design of robust, mutation-resistant inhibitors. Intervariant differences are also observed among the dimer interfaces, which is another potential inhibitor target.

MATERIALS AND METHODS

Sequence Analysis and Clustering. SARS-CoV-2 genome sequences were found by searching the GISAID (https://www.gisaid.org/)[28] EpiCoV database on May 3, 2020, using the host keyword "human" and a submission cutoff date of April 29, 2020, yielding a total of 15 432 SARS-CoV-2 genomes. Genomes outside the range of ±3% reference (RefSeq: NC 045512.2) length (29 006 bp−30 800 bp inclusive) or ≥1% N content were removed, leaving 10 644 "high-quality" sequences. Open reading frames in these high-quality full genomes were compared with a reference M^{pro} nucleotide sequence (WT, RefSeq, NC 045512.2; loc, 10 055−10 972) to extract M^{pro} sequences of at least 80% similarity using a script written in Python v3.7.0.[29] Genomes with gaps or ambiguous nucleotides (e.g., N, S, D, per International Union of Pure and Applied Chemistry (IUPAC) nomenclature[30]) in the M^{pro} sequence were excluded from this data set, leaving a total of 10 578 sequences from high-quality genomes.

Nucleotide sequences were converted into amino acid sequences and screened for nonsynonymous mutations against the WT M^{pro} using code written in Wolfram Mathematica 12.1,[31] yielding 511 nonsynonymous mutations in M^{pro}, 77 of which were unique. A single unique M^{pro} variant, found in an April 24, 2020 data set, but no longer available in the GISAID database, was also used in our analyses. M^{pro} variants were used in phylogenetic analyses along with reference human, bat, and pangolin viruses. Full genome alignments were performed using MUSCLE (v3.8.1551, max 8 iterations, enable find diagonals)[32] on the complete set of nonsynonymous M^{pro} mutants as well as reference WT, bat, and pangolin sequences. Trees were generated in MEGA X,[33] using the Neighbor-Joining method;[34] a bootstrap test[35] of 1000 replicates was performed, and distances were calculated using the Maximum Composite Likelihood model.[35] In all, 515 full genomes were used in phylogenetic analyses; 78 unique M^{pro} mutants and a reference WT sequence (79 total) were used for molecular modeling.

Molecular Modeling of Wild-Type and Variant Protein Structures. Initial conditions for the WT trajectories used here are based on the PDB structure 6Y2E,[36] representing a mature (i.e., cleaved pro-sequence) protein. For monomer trajectories, the A chain of 6Y2E was employed. Initial variant monomer and dimer structures were, respectively, predicted using MODELLER 9.23,[37] using the 6Y2E structure as a template; three rounds of annealing and MD refinement were performed using the "slow" optimization level for each (final objective function values are provided in Table S7). Initial structures were then processed to correct protonation states to reflect their predicted cellular environment (with protonation states predicted using PROPKA 3.1[38]). Each corrected model structure was then minimized and equilibrated in explicit solvent; simulations were performed using NAMD[39] with the CHARMM36 force field[40] in TIP3P water[41] at 310 K under periodic boundary conditions (with a 10 Å margin water box). Solvated protein models were energy-minimized for 10 000 iterations before being simulated for 0.5 ns, followed by a water box size adjustment, after which a 10 ns trajectory was simulated with conformations being sampled every 20 ps; an *NpT* ensemble was used, with temperature controlled via Langevin dynamics with a damping coefficient of 1/ps and Nosé−Hoover Langevin piston pressure control set to 1 atm.[42,43] Final conformations from each trajectory were used

https://dx.doi.org/10.1021/acs.biochem.0c00462
Biochemistry XXXX, XXX, XXX−XXX

to generate figures in the main text and Supporting Information.

Network Analysis. A protein structure network (PSN) was calculated for each modeled conformation of each variant via scripts employing the statnet,[44−46] Rpdb,[47] and bio3d[48] libraries for R.[49] Vertices were defined using the method used in ref 50, where each node represents a chemical moiety, with edges being defined by interatomic contacts. Specifically, two nodes i and j are considered adjacent if i contains atom g and j contains atom h such that the g and h distance is less than 1.1 times the sum of their respective van der Waals radii (using values from ref 51). The node definitions are illustrated in Figure S1A, and a small-moiety PSN of this type for WT M^{pro} is shown in Figure S1B. Active site networks (ASNs) were constructed from each PSN as described in ref 52. Briefly, all vertices belonging to the catalytic Cys and His residues were identified, along with all vertices adjacent to these vertices within the PSN. The ASN was then defined as the subgraph of the corresponding PSN induced by this combined vertex set (Figure S1C). In the case of dimer models, one ASN was constructed for each chain.

To assess overall cohesion, degree k-core values[53] were calculated for each vertex in each PSN, and the average core number was computed for the entire protein and for the vertices in each domain, respectively. All calculations were performed using the sna library[46] for R. For dimer PSNs, interfacial moieties were identified by selecting all vertices adjacent to at least one vertex in the opposing chain, and average core numbers were computed for these interfacial vertices to assess cohesion in the dimerization interface; raw counts of edges spanning the two chains (interfacial tie volumes) were also calculated. For each vertex associated with a moiety in the active site, three measures identified as associated with active site constraint by ref 52 were computed: the degree, or number of ties to other vertices; the triangle degree, or number of triangles (3-cliques) to which the vertex belongs; and core number, or number of the highest degree k-core[54] to which the vertex belongs. Physically, these, respectively, indicate the total number of contacts associated with the chemical group (potentially impeding its motion), the number of truss-like, triangular structures in which the group is embedded (again, restricting mobility), and the extent of local cohesion around the chemical group, which is found to distinguish "tighter" and "looser" packing regimes.[55] To summarize the impact of each measure over the active site as a whole, values were averaged across active site vertices. As an additional constraint measure, the number of paths between each pair of active site vertices through neighboring (i.e., nonactive site) vertices was computed, and the log of the minimum of this value over the set of active site vertex pairs was employed as a measure of site cohesion. Intuitively, high values of site cohesion indicate that all active site chemical groups are connected by a large number of indirect contacts, while low values suggest that at least one pair of active site moieties has few local pathways holding them together. These four indices (mean active site degree, mean active site triangle degree, mean active site core number, and site cohesion) were used to produce an omnibus index of site constraint via principal component analysis (PCA) of the standardized network measures over all modeled conformations, as described in ref 52. This first principal component (the constraint score) accounted for approximately 71% of the variance in all four measures in the case of monomer ASNs, and the ratio of its associated eigenvalue to the next largest was approximately 4.7 (confirming the dominance of the principal eigenvector). This process was repeated for the dimer ASNs, resulting in a constraint score vector accounting for approximately 69% of the total index variance, with an eigenvalue ratio of approximately 4.3.

Comparing Mean Cohesion and Constraint Scores Across Variants. Because cohesion and constraint scores are heavily autocorrelated within trajectories, we employ a parametric bootstrap strategy to obtain autocorrelation-corrected standard errors and confidence intervals.[56] For each time series of scores for each trajectory, an autoregressive (AR) model with AIC-selected order was fit, and the estimated series mean was obtained (estimation performed by maximum likelihood estimation using the ar function in R[49]). The whitened residuals from the time series model were then used to construct 5000 parametric bootstrap replicate series, which were then refit to obtain bootstrap replicate means. Mean estimates from the bootstrap replicates were used to construct 95% bootstrap confidence intervals and standard errors for the series mean. This procedure was applied to the MD trajectory for each variant. For cohesion scores, mean and bootstrap standard errors are provided for the full protein and each domain in Table S3.

Kernel PCA of Active Site Networks. To identify key features differentiating active site conformations observed throughout the entire sample of trajectories, we employ a kernelized principal component analysis (kernel PCA[57]) of a stratified sample of monomer and dimer ASNs from all M^{pro} variants. The ASN sample consists of 25 evenly spaced conformations from each trajectory (monomer, dimer A chain, and dimer B chain), for a total of 5925 networks. For comparative purposes, each ASN was mapped to the set of all unique vertices appearing in any ASN in the sample, and the upper triangle of the associated adjacency matrix was vectorized, yielding a binary vector of fixed dimension encoding each network. Analysis of the vectorized ASNs was performed using a disjunctive normal form (DNF) kernel,[58] defined by

$$k(x, y) = (1 + \epsilon)^{x^T y} - 1$$

where x and y are respective ASN vectors, and ϵ is a free parameter controlling regularization. In a graph-theoretic context, the feature space of the DNF kernel consists of the set of all labeled subgraphs of the inputs; thus, kernel PCA on this space identifies combinations of labeled subgraphs that efficiently discriminate among ASNs. Hyperparameter tuning was performed via grid search, with the performance assessed by reconstruction of a held-out sample of 100 randomly selected ASNs, following training on a separate sample of 1000 ASNs. Reconstruction was performed using the preimage method from ref 59 with a local sample of 10 neighbors; the optimal hyperparameter obtained was $\epsilon = 0.0739$. Following hyperparameter tuning, principal component scores were calculated for use in subsequent analysis. Analyses were performed using R[49] with portions implemented using Rcpp.[60]

RESULTS AND DISCUSSION

Mutations in M^{pro} Are Geographically Distributed. From the GISAID (https://www.gisaid.org/)[28] EpiCoV database (through April 29, 2020), 78 unique nonsynonymous mutations to M^{pro} were found in addition to the WT sequence,

https://dx.doi.org/10.1021/acs.biochem.0c00462
Biochemistry XXXX, XXX, XXX−XXX

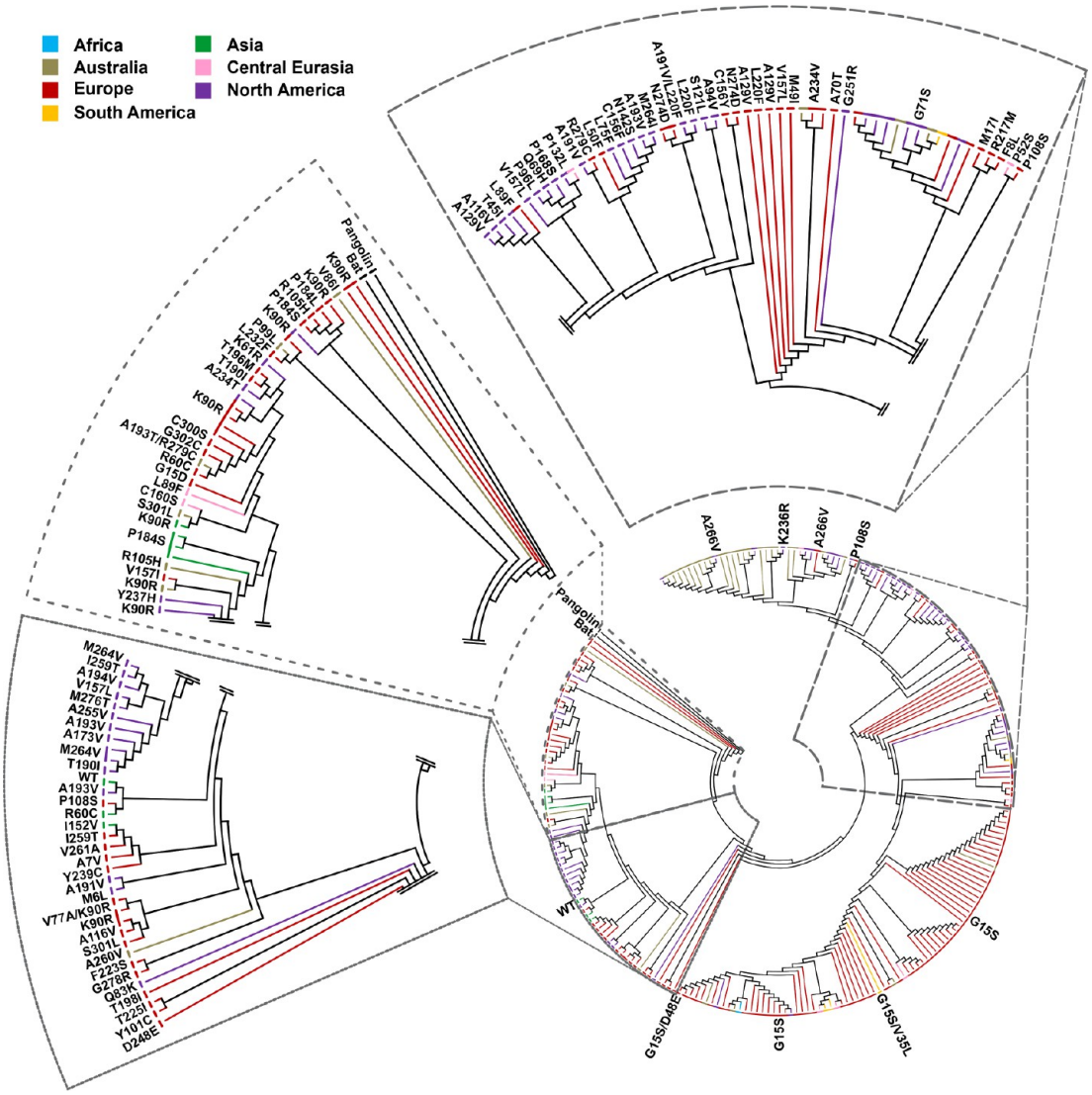


Figure 1. Optimal tree generated using 512 full mutant genomes and three reference genomes: human wild-type (WT),[64] bat,[3] and pangolin.[65] Only topology is shown; branch lengths are not to scale (average branch length = 1.432161×10^{-4} base substitutions per site). Each continuous arc corresponds to a variant label; these represent only adjacent branches with the same mutation in M^{pro} and do not necessarily indicate shared ancestry. Branches and arcs from human clinical samples are color coded by location, which includes the following subregions: Africa, light blue (Democratic Republic of the Congo); Asia, green (Beijing, Fujian, Malaysia, Shanghai, Vietnam, and Wuhan); Australia, gold; Central Eurasia, pink (Georgia, Jordan, Russia, and Turkey); Europe, red (Belgium, Denmark, England, Finland, France, Germany, Iceland, Luxembourg, Netherlands, Scotland, Spain, Sweden, Switzerland, and Wales); North America, purple (Costa Rica and United States of America); South America, yellow (Argentina and Brazil). Subtrees that contained identical subregions and mutations have been condensed into a single branch; all subtrees and their constituent accessions can be found in Table S1.

including 73 single point variants and 5 double variants. For genome sequences containing these M^{pro} variants, full genome alignments were performed using MUSCLE,[32] and neighbor-joining trees were generated using MEGA X.[33] Overall, the variation in SARS-CoV-2 sequences observed so far is relatively low, with mutation hotspots not evenly distributed throughout the genome, but localized to specific sequence regions.[61] Because M^{pro} is critical for viral replication, mutations that have a large deleterious effect on virus replication are unlikely to be observed in clinical isolates; all M^{pro} variants investigated here are therefore assumed to be enzymatically competent. In general, codon usage and amino acid frequency in viruses of eukaryotes are essentially identical to those of their eukaryotic hosts, reflecting the viruses' use of the host translation machinery.[62]

The known mutations in M^{pro} are summarized in Figure 1. The tree was generated based on overall genome similarity; however, only sequences containing at least one non-synonymous mutation in M^{pro} were included in the analysis, along with the WT human sequence and two nonhuman reference sequences. The accession numbers and geographical sources are listed in Table S1. The solid arcs around the outside of the diagram indicate M^{pro} mutations; color coding corresponds to the geographical source. Several mutations appear to have arisen more than once in the virus's evolutionary history so far. Notably, K90R variants appear in multiple distantly related subtrees; five of these unique evolutionary events can be verified in Nextstrain's SARS-CoV-2 phylogenetic tree.[63] Further, L89F, P108S, and N274D arise at least twice in both trees.

https://dx.doi.org/10.1021/acs.biochem.0c00462
Biochemistry XXXX, XXX, XXX−XXX

These phylogenetic comparisons appear to support a multiple event hypothesis but are subject to errors resulting from the sparsity of testing. The repeated occurrence of the same mutation in seemingly unrelated subtrees may be due to missing data that would show their evolutionary connectedness. The average branch length of Figure 1, which shows only topology, is 1.432161×10^{-4} base substitutions per site (including those from the bat[3] and pangolin[65]); 32.2% of the 1028 branches have, to ten significant figures, 0 base substitutions per site. For a genome of roughly 30 000 base pairs, this amounts to an average of only 4 substitutions per branch. All of these unique mutants therefore effectively belong to the same strain, making them difficult to place in an evolutionary context. For more diverged mutants, unfortunately placed ambiguous nucleotides[30] could push them from one subtree to another. With the exception of five double variants, the sequences in Figure 1 arise from single point mutations. Whether and how M^{pro} mutations have affected viral fitness is not yet known, but at least three mutants have remained in the population long enough to accumulate another mutation: L220F to A191V/L220F, G15S to G15S/D48E and G15S/V35L, and K90R to V77A/K90R. It is worth noting, that although a single variant A191V exists, the A191V/L220F double variant likely stemmed from an L220F ancestor due to its shared lineage with L220F single variants. A fifth double variant, A193T/R279C, was found but did not stem from any single mutation in our data set; its origins remain unclear.

Although a mutation's prevalence and evolution in a population may be interpreted as a sign of robust viral function, the opposite does not necessarily indicate reduced infectivity. Testing rates, social behavior, and time of first infection in each region are all factors that contribute to the spread of the disease and the availability of sequencing data. For instance, a large number of K90R mutants were collected in Iceland, where the number of tests per 1000 people is nearly twice as many as the next leading country's and more than seven times as many as the United States' (Iceland, 141.75; USA, 18.21, as of April 29, 2020).[66] Consequently, further investigation is needed to determine whether M^{pro} mutations affect viral fitness on a global scale. As such, without greater divergence and more sequences, it is difficult to tell if the presence of an M^{pro} mutation in unrelated subtrees is evidence of multiple evolutionary events, or an artifact of sparse testing. Despite these caveats, it is clear that these 78 variants are functional enough to infect people within their local populations. In the structural analyses that follow, we focus on the differences in protein properties of the clinically observed M^{pro} variants relative to WT.

Because only sequences harboring M^{pro} mutations were retained for analysis, certain geographical areas appear to be underrepresented. It is likely that the strains that had spread to underrepresented regions prior to our data collection simply did not have M^{pro} mutations. Different regions tend to be dominated by different mutants, a feature that might be explained by the timing at which these mutations arose or arrived. For instance, 83 of the 100 M^{pro} mutants from Iceland were K90R, and most stemmed from a single shared ancestor (see Supporting Information). Further, it is likely that heterogeneity in sequencing rates have resulted in a less-than-complete data set. As of April 29th, the only North American, South American, and African M^{pro} mutants reported in the GISAID database that passed our filtering parameters were from Costa Rica and the USA, Argentina and Brazil, and the DRC respectively. This does not necessarily indicate a lack of M^{pro} mutations in other subregions and may instead reflect differences in sequencing rates.

It is worth noting that some nodes in expanded Figure 1 (see Supporting Information) exhibit bootstrap values less than 10 (i.e., their branch relationships were shown in fewer than 10% of replicates). The majority of these extremely low values can be found on large subtrees that share an M^{pro} mutation, like D248E or K90R, many of which likely spread from a common source. Their branches can often be interchanged with no reduction in accuracy due to high sequence identity. Low bootstrap values here primarily reflect genotype similarity within subtrees rather than low accuracy. This is a known property of the Felsenstein procedure,[35] which bootstraps loci instead of sequences; low bootstrap values indicate groupings that are sensitively dependent upon differences in small numbers of loci but do not necessarily reflect uncertainty due to sequence sampling.

M^{pro} Mutations to Date Suggest Selection for Larger, More Massive, and More Hydrophobic Residues. To reveal the global pattern of substitutions, we visualize mutations in M^{pro} independent of sequence position or location in the three-dimensional structure, by a network where the nodes, or vertices, are amino acid types and the edges (represented by arrows pointing in the direction of substitution) are directional indicators of how often one amino acid was observed to substitute for another (Figure 2). The weights of the edges indicate the frequency of the mutation across known M^{pro} variants, while node color reflects residue hydrophobicity on the scale of Kyte and Doolittle.[67] The most obvious trend observed in the pattern of mutation so far is the preferential substitution of larger, more hydrophobic amino acids in place of smaller, less hydrophobic ones. The pattern is consistent with increased incidence of amino acid types that

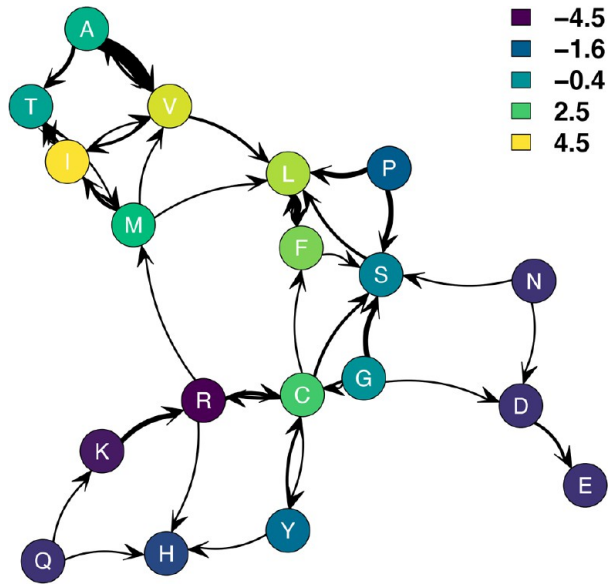


Figure 2. Amino acid substitutions in SARS-CoV-2 M^{pro} observed up to April 29, 2020. Arrows indicate the direction of substitution: an arrow from *i* to *j* indicates at least one clinically observed substitution of residue type *i* to residue type *j*; heavier lines indicate larger numbers of observed substitutions. Color indicates hydrophobicity, using the scale of Kyte and Doolittle.[67] In general, substitution has been toward larger and more hydrophobic residues.

https://dx.doi.org/10.1021/acs.biochem.0c00462
Biochemistry XXXX, XXX, XXX−XXX

are more likely to be present in folded domains rather than those found more often in less-structured linker regions.[68]

In particular, it is notable that alanine has very few incoming ties and a large number of outgoing ties, mostly to valine, which has a larger and more hydrophobic side chain. Alanine is at the same time one of the most common amino acids and one of those with the most variable prevalence in the human genome.[69] Similarly, observed ties to isoleucine are mostly incoming from smaller residues; leucine likewise has more incoming than outgoing ties, with the bulk of its outgoing ties going to phenylalanine, which is also large and hydrophobic. However, aromatic residues per se do not appear to be selected. For example, tyrosine has mostly outgoing ties. Also notable are the selection away from the secondary structure breakers proline and glycine, both of which have only outgoing ties, and the propensity for lysine to be replaced by arginine even though both side chains are positively charged. Arginine is both larger and capable of making more and stronger hydrogen bonds, as well as cation-π interactions not available to lysine, leading to its known overrepresentation in interdomain and intermonomer interfaces.[70−73]

The mean differences in side chain properties for observed M^{pro} mutations are summarized in Table 1. As observed in the network representation (Figure 2), on average, the mutated residues are larger and more hydrophobic than the ones they replace. Although substituted residues are on average larger and more massive, we do not see strong evidence favoring bulky residues over more compact ones, independent of mass: residue bulk (measured as volume/mass) for substituted residues did not differ significantly from WT (mean difference = 0.02 Å^3/Da, t = 1.87, p = 0.0650). The variant sequences are not significantly different from WT in charge or aromatic content.

Table 1. Mean Differences in Side Chain Properties for Substituted Residues versus WT (N = 83; Substitutions from Double Mutants Considered Separately)[a]

	mean difference	std error	t value	p value
polar (1 = true)	0.08	0.07	1.22	0.2251
hydrophobicity	1.03	0.30	3.47	0.0008[d]
charge	−0.05	0.04	−1.27	0.2078
aromatic (1 = true)	0.07	0.04	1.62	0.1093
mass (Da)	9.97	3.49	2.85	0.0055[c]
volume (Å^3)	11.58	3.65	3.17	0.0021[c]
bulk (Å^3/Da)	0.02	0.01	1.87	0.0650

[a]On average, substituted residues are significantly more hydrophobic, more massive, and larger in volume than those they replace (all p values for two-tailed t-tests versus no difference). [b]$P < 0.05$ [c]$P < 0.01$ [d]$P < 0.001$

M^{pro} Mutations Occur Throughout the Protein. In order to provide context for the discussion of conformational changes and the locations of mutation sites, a ribbon diagram of the SARS-CoV-2 M^{pro} monomer is shown in Figure 3A. The three domains, the active site residues, and other structural features are labeled. The dimer is shown in Figure 3B; the dashed box shows the approximate location of the interchain salt bridge shown in Figure 3C. In SARS-CoV-1 M^{pro}, this salt bridge between Arg4 of Chain A and Glu290 of Chain B is an important component of the dimer interface,[74] and substitution of either or both of these residues with alanine resulted in shifting the equilibrium from dimer to monomer, although dimerization was not abolished completely.[14] Panels D and E show two different views of the dimer interface, with key residues labeled. In addition to important interfacial residues, we also highlight key residues near the oxyanion hole, a structural feature found in many serine and cysteine proteases that stabilizes negatively charged, tetrahedral transition states.[75]

In SARS-CoV-2 M^{pro}, several interactions at or near the dimer interface play a role in keeping the oxyanion hole in the active conformation. An intrachain hydrogen bond between the side chain NH_2 of Arg298 and the backbone carbonyl of Met6 appears to help hold the N-terminus (also called the N-finger) in place to make the appropriate contacts with the other chain.[76] In the S1 specificity pocket, a ring-stacking interaction between the side chains of His163 and Phe140 is stabilized by a network of hydrogen bonds comprising residues from both monomers. Hydrogen bonding interactions between Glu166 and Ser139 of Chain A with the N-terminus of Chain B (Ser1) are also important for stabilizing the dimer interface, along with interactions between the Ser10 residues of both chains. Mutating the adjacent Gly11 to Ala completely abolishes dimerization due to a dramatic conformational change to the N-finger that causes disruption of the interface.[21] Although the mutational studies discussed here were performed using SARS-CoV-1 M^{pro}, many of the key residues are also found in SARS-CoV-2 M^{pro}, and we hypothesize that they play similar roles.

Figure 4A shows the M^{pro} dimer with mutation sites indicated by spheres. The mutations observed so far are relatively evenly distributed throughout the protein. Notably, mutations are tolerated close to the active site residues, near the oxyanion hole, and at the termini, both of which are involved in modulating the dimer conformation.[17] Two different views of the dimer interface are shown in Figure 4B, with interfacial residues shown as space-filling models and with mutated residues highlighted in light blue (Chain A) or dark blue (Chain B). Despite the importance of the dimer interface, several mutations are observed in this region: M6L, A7V, A116V, S121L, C300S, S301L, and G302C. Particularly notable are M6L and A7V, which are part of the N-finger that participates in interactions with the C-terminal domain of the second monomer. In both cases, the hydrophobic character of the residue is preserved, with a slight increase in the size of the residue. With the exception of A116V, the other mutations represent qualitative changes in side chain properties. These changes in chemical properties are apparently well tolerated, as all of these mutations were found in clinical isolates.

Molecular Modeling Suggests Regionally Specific Differences in M^{pro} Variant Structure. For WT M^{pro} and each variant, molecular models of the monomer and dimer structures were constructed using MODELLER 9.23,[37] based on PDB structure 6Y2E,[36] followed by annealing, correction of protonation states, and all-atom molecular dynamics simulation in explicit solvent (see Materials and Methods). Examples of representative models are shown in Figure S2, intramolecular contacts for residue 225 in WT and the T225I variant are shown in Figure S3 and listed in Table S2, and selected double mutants are shown in Figure S4. The positions of all observed K to R and R to C mutations, which are commonly observed in interfaces, are shown in Figure S5. In M^{pro}, two such mutations, K236R and R279C, are located in domain III, near the dimer interface. The positions of all mutated residues are mapped onto the WT structure in Figure

https://dx.doi.org/10.1021/acs.biochem.0c00462
Biochemistry XXXX, XXX, XXX−XXX

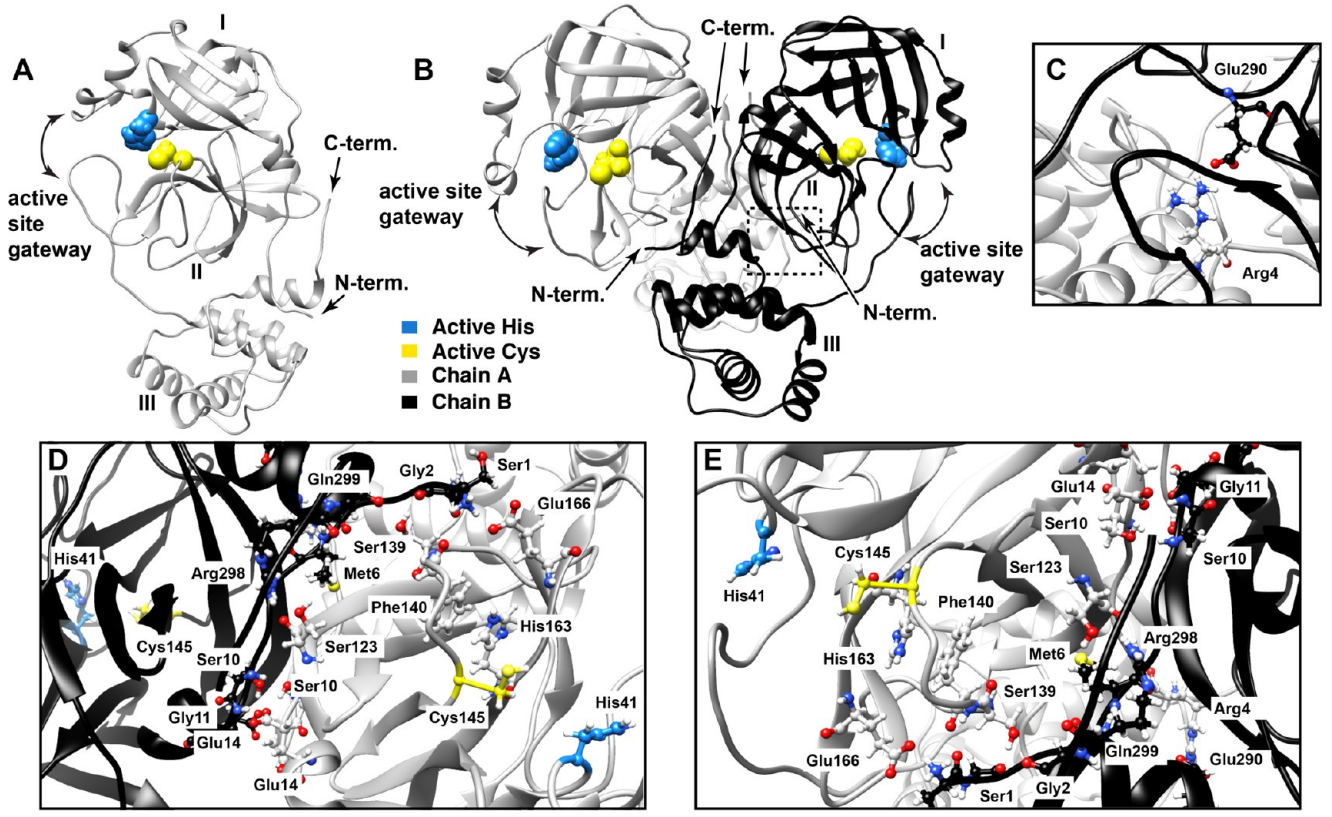


Figure 3. (A) Ribbon diagram for the M^{pro} monomer, with key structural features labeled, including the three domains, the N- and C-termini, and the active site residues, H41 and C145. (B) Ribbon diagram for the M^{pro} dimer, with separate chains shown in light gray and black. The N- and C-termini, both of which mediate interactions at the dimer interface, are labeled for each monomer. The dashed square shows the approximate location of the salt bridge shown in panel C. (C) A salt bridge between Arg4 of one monomer and Glu290 of the other is shown. This salt bridge is hypothesized to be important for stabilizing the dimer interface and maintaining the active conformation. (D, E) The dimer interface is shown from two different angles, with key residues labeled.

S6, with color coding indicating the change (if any) in side chain chemical properties. We do not observe gross differences in structure or dynamics across variants, as expected given that all variants were found in clinical isolates and are therefore necessarily functional; mutations leading to radically altered or misfolded structures would likely be strongly selected against. However, analysis of MD trajectories does suggest more subtle differences across variants, providing insight into function-preserving changes.

To assess the overall degree to which local structure is conserved across M^{pro} variants, we compute the cross-variant variance in average ϕ,ψ backbone torsion angles by residue within free monomers. In order to control for overall flexibility, we normalize this by the estimated variance in torsion angles within each trajectory. For arbitrary angle α_i at residue i, this leads to the local variation index

$$v(\alpha_i) = \log \frac{\mathrm{Var_B}(\overline{\alpha}_i)}{\frac{1}{N}\sum_{j=1}^{N} \mathrm{Var_W}(\alpha_{ij})}$$

where α_{ij} is the vector of angles of type α_i over the trajectory of variant j with corresponding angular mean $\overline{\alpha}_{ij}$, $\overline{\alpha}_i$ is the vector of such means across variants, $\mathrm{Var_B}$ is the "between variant" angular variance in mean angles, and $\mathrm{Var_W}$ is the "within variant" angular variance in α_{ij}. Intuitively, high values of $v(\alpha_i)$ indicate relatively large between-variant variation in α_i relative to angular variation seen within the trajectories themselves. For $v(\phi_i)$ and $v(\psi_i)$, such values correspond to systematic changes in local conformation associated with M^{pro} mutations. By turns, low values of $v(\phi_i)$ and $v(\psi_i)$ indicate residues whose local structure does not vary meaningfully across variants. It should be noted that such regions can be either flexible or rigid.

Figure 5 shows the mean local variation indices for ϕ,ψ by the residue for the 79 M^{pro} variants, indicated by color on the structure of M^{pro} WT. (Separate values for ϕ and ψ are shown in Figure S7.) It is immediately noteworthy that, with the minor exception of two small loop regions around N277 and F223 (respectively), Domain III shows little systematic variation across variants. The β-sheet-rich structure around the active site is also relatively well-conserved. By contrast, we see relatively high levels of between-variant difference in the interdomain region involving the termini (residues G2-A7 and S301−F305) and the double loop "active site gateway" region involving (respectively) L50-Y54 and D187-A191. The former is potentially significant in influencing large-scale flexibility (possibly relevant to dimerization), whereas the latter is of obvious relevance to substrate processing and specificity. This motivates a more detailed examination of variation in the active site, to which we return below.

The relatively high levels of conformational variation in the interdomain regions suggest functionally relevant differences in global cohesion across variants. To assess this, we employ protein structure networks (PSNs), which are well-suited for assessing the looseness or cohesiveness of contacts among chemical groups.

https://dx.doi.org/10.1021/acs.biochem.0c00462
Biochemistry XXXX, XXX, XXX−XXX

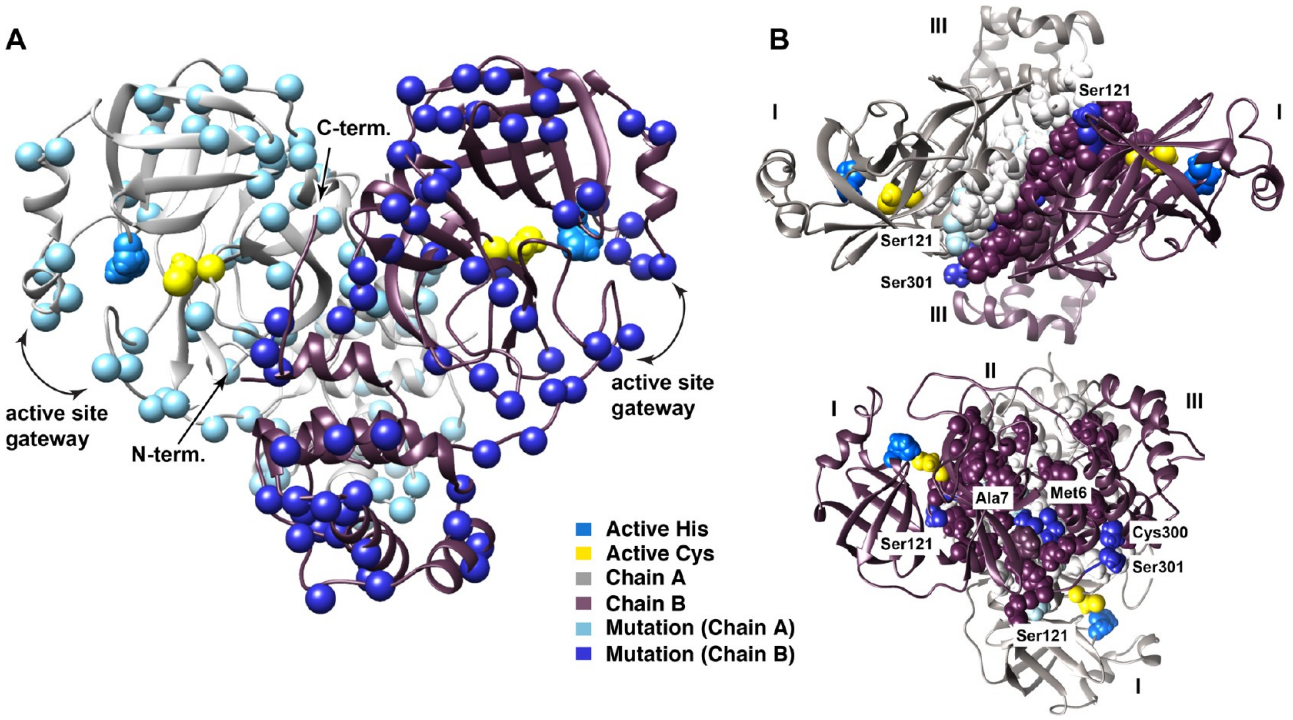


Figure 4. (A) Ribbon diagram for the M^{pro} dimer, with separate chains shown in light gray/light blue and purple/dark blue. The locations of all observed mutations are indicated by spheres centered on the α-carbons. (B) Two different views of the dimer are shown, with all residues at the dimer interface rendered as space-filling models, providing a visualization of the size of the dimer interface and its spatial relationship to the active site residues (blue, His41; yellow, Cys145) Residues having at least one mutation in this data set are highlighted in light blue (Chain A) or dark blue (Chain B.).

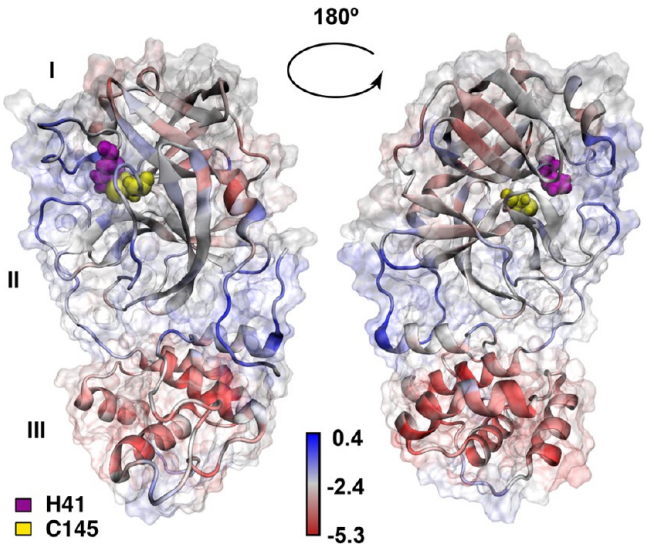


Figure 5. Local variation indices for M^{pro} monomer backbone torsion angles (front/back views). Blue residues show higher levels of cross-variant ϕ,ψ differences relative to baseline variation; red residues show little evidence of structural difference across variants. Domain III is substantially conserved, while greater change is seen in the interdomain regions and loop regions adjacent to the active site.

For all M^{pro} variants, moiety-level PSNs were constructed for each frame within each variant trajectory, using the definitions found in ref 50 (Figure S1) The assessment of global cohesion was performed by computing the mean degree k-core number for all moieties in each structure; to allow comparison of global cohesion within domains, we also compute mean core numbers within each of the three domains. The mean core number can be considered an index of structural cohesion, with higher values indicating greater numbers of redundant contacts among chemical groups.[55] To account for within-trajectory autocorrelation in comparing mean core numbers, autocorrelation-corrected parametric bootstrap confidence intervals and standard errors were employed.

Figure 6 shows global and domain-specific cohesion levels (i.e., mean core numbers) for all variant monomers, sorted in descending order of mean cohesion. Means and standard errors for each variant can be found in Table S3. As suggested from the torsion angle analysis, cohesion differs significantly among variants, both globally and within domains. On average, the majority of variants are estimated to be less cohesively structured than WT, with the exception of Domain III (in which WT does not differ significantly from the mean). It is possible that these differences indicate selection for more globally flexible structures (again, with the exception of Domain III). Whether or not this is the case, however, it appears clear that less cohesive structures are not strongly selected against. Such flexibility may affect dimerization kinetics, which is relevant to protease function and the development of robust dimerization inhibitors.

While cohesion and flexibility within M^{pro} monomers may provide clues to how mutations may impact dimerization, examination of the interfacial region within M^{pro} dimers suggests indications of the impact of mutations on the behavior and stability of the dimers themselves. Figure 7A shows mean cohesion scores (i.e., mean degree k-core numbers) for interfacial moieties in dimer trajectories for all M^{pro} variants, together with autocorrelation-corrected 95% bootstrap confidence intervals; values are also provided in Table S4. (We here define a moiety to be interfacial if it is adjacent to at least

https://dx.doi.org/10.1021/acs.biochem.0c00462
Biochemistry XXXX, XXX, XXX−XXX

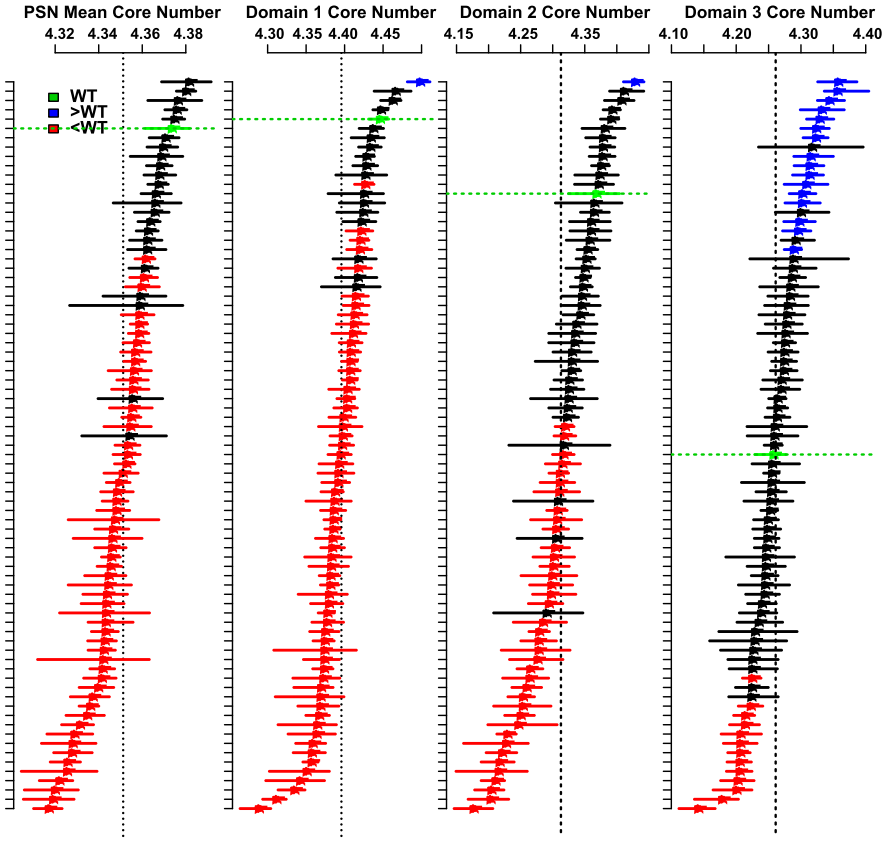


Figure 6. Mean core numbers for M^{pro} monomer PSNs, by variant (ordering is by mean value in each panel). Points indicate trajectory means, with segments showing autocorrelation-corrected 95% bootstrap confidence intervals; red/blue intervals have t values versus WT (green) of at least ±2, indicating significant variation in structural cohesion across variants. Overall, the majority of variants are less cohesive than WT globally and in Domains I and II, while Domain III cohesion in WT is typical of the variant set.

one chemical group in the opposite chain (Figure 7C). This is assessed on a frame-by-frame basis, allowing us to account for changes in contact patterns within trajectories.) Although there is some variation in interfacial cohesion, relatively few trajectories differ significantly from the mean cohesion level, and very few are two standard errors away from wild-type (12 being more cohesive and 1 being less cohesive). This is consistent with the hypothesis that interfacial cohesion is strongly conserved in M^{pro}, suggesting that a looser or more dynamic dimer core may impede function. (The fact that, of those variants differing significantly from WT, all but one show greater cohesion is also consistent with this determination, albeit not determinative.)

For a different look at interaction across the interfacial region, we also consider the total number of contacts between the A and B chains over time (the tie volume across the dimer interface, e.g., the interfacial ties in Figure 7D). Figure 7B shows mean dimer interface tie volumes and associated confidence intervals for the M^{pro} variants, with variants sorted in the order of the mean core number to facilitate comparison (panel A). Whereas the mean core number provides a measurement of overall structural cohesion at the interface (including both interactions within and between chains), the tie volume directly measures the number of cross-chain contacts and is thus a potential proxy for interaction strength. As with cohesion, tie volume is fairly well-conserved, and few trajectories show significant deviation from wild-type (with seven trajectories having higher tie volume than WT, and two lower). Interestingly, interfacial tie volume is not strongly related to interfacial cohesion, indicating that the tightness or looseness of the structure around the interface is not simply a function of the raw number of interfacial contacts. Both, however, suggest a relatively high level of conservation in interfacial structure in the M^{pro} dimer across variants observed to date. While the first version of this work, which addressed moiety-level networks of M^{pro} monomers[77] was under review, a preprint[78] was released, showing similar results using residue-level network analysis on MD trajectories of M^{pro} dimers, and subsequently published.[79] Although different methodology is used, the conclusions are in broad agreement, with significant flexibility observed in the N-finger region that is important for dimerization, as well as the loops near the active site. The authors also describe key low-frequency motions of the dimer that may be impacted by mutation and identify a pocket near the dimer interface that could serve as a potential target for allosteric inhibitors, underscoring the importance of mutational analysis as a part of effective inhibitor design. Moiety-level network analysis of M^{pro} dimers was then added to this paper in response to reviewers' comments, again finding substantial agreement with the residue-level results.[78]

Active Site Networks Suggest Potential Activity Differences across M^{pro} Variants. The observation of structural variation in loop regions associated with the binding pocket motivates closer examination of variation in the M^{pro} active site. To this end, subgraphs of the full protein structure networks comprising moieties belonging to the active site residues and their neighbors were constructed to produce active site networks (ASNs)[52] for all conformations. A

https://dx.doi.org/10.1021/acs.biochem.0c00462
Biochemistry XXXX, XXX, XXX−XXX

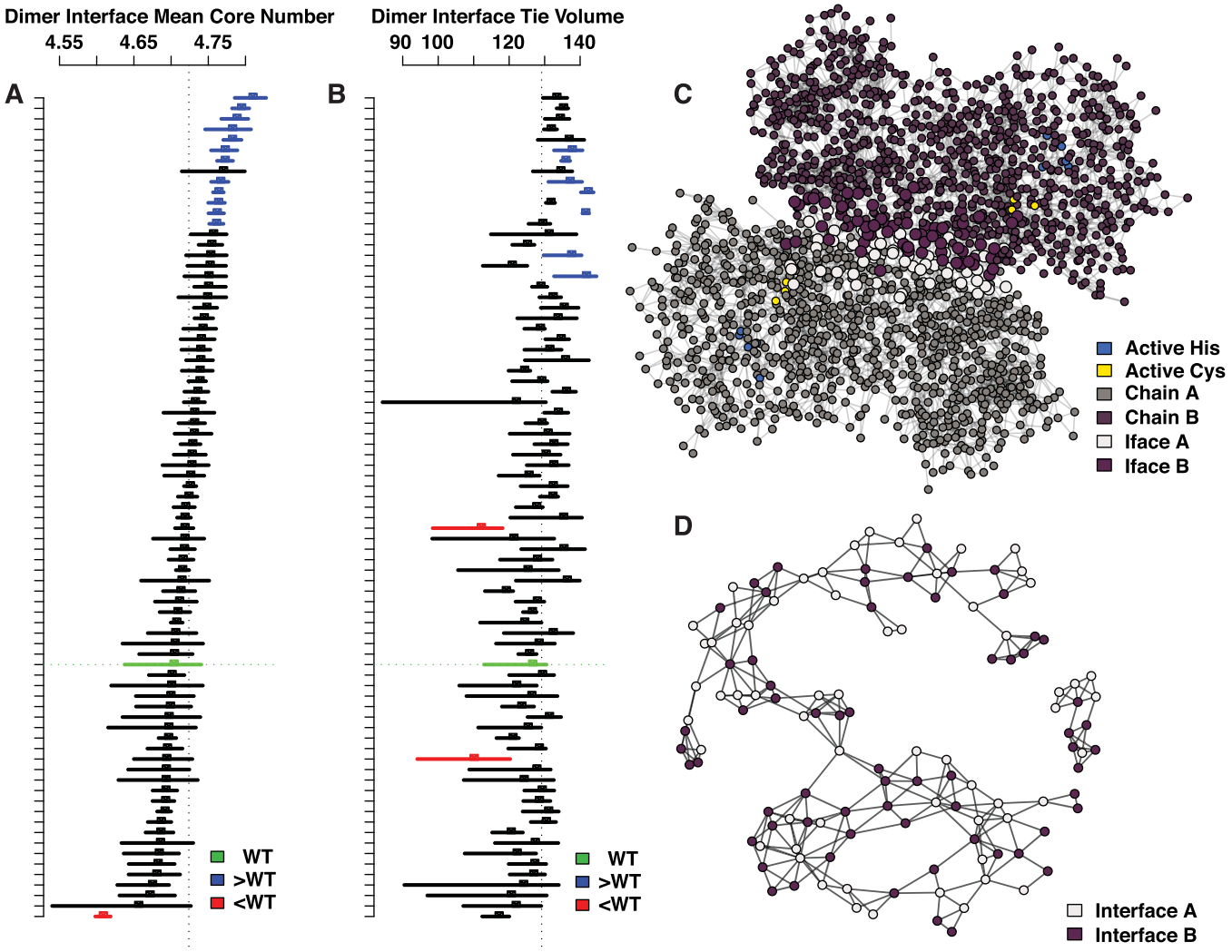


Figure 7. (A) Mean core number for moieties at the dimer interface across all variants, with 95% autocorelation-corected bootstrap confidence intervals. Higher values indicate a higher core number (greater cohesion); variants indicated in red/blue have significant variation in core number relative to WT (green). The vertical dotted line indicates grand mean. (B) Mean tie volume across the dimer interface, ordered per panel A. Variants indicated in red/blue have significantly lower/higher tie volumes across the interface relative to WT (green). (C) Illustrative moiety-level PSN for the WT M^{pro} dimer. Moieties associated with chains A and B are indicated in gray and dark purple, respectively. Moieties making up the interfacial residues for chains A and B are indicated in light gray and light purple, respectively. Also shown are moieties associated with the active site His (blue) and Cys (yellow.) (D) Induced subgraph of the PSN in panel C based on interfacial moieties. Interfacial tie volume is the count of ties from A (gray) to B (purple) interfacial nodes, while embeddedness in locally cohesive structures contributes to the mean *k*-core number.

protein's ASN describes physical interactions among active site moieties and other groups that are immediately adjacent in the 3D structure, irrespective of their positions in the amino acid sequence. Per ref 52, we compute for each ASN a constraint score, a general measure of active site flexibility that is associated with substrate specificity. The constraint score is the first principal component of a set of several network metrics (see Materials and Methods), with higher values indicating a greater tendency for the catalytic residues to be constrained by cohesive contacts with other residues, and lower values indicating fewer such constraints. Examples of ASNs corresponding to the maximum, minimum, and mean observed constraint values over all observed M^{pro} monomer conformations are shown in Figure 8. For the dimer structures, we repeat this process for the active site in each chain, yielding two constraint values per variant; these scores, along with corresponding maximum, mean, and minimum constraint ASNs, are shown in Figure 9. Values for both monomers and dimers are also provided in Tables S5 and S6.

Examination of the mean constraint scores for each variant trajectory suggests potential activity differences across M^{pro} variants. Figure 8A shows mean constraint scores for each variant monomer, with autocorrelation-corrected parametric bootstrap confidence intervals. Of the 79 trajectories examined, 22 (28%) were significantly below the grand mean (dotted vertical line) and 28 (35%) were significantly above it; similarly, when directly compared to WT, 12 variants were observed to be significantly less constrained, while 17 were significantly more constrained (i.e., bootstrap *t*-scores less than −2 or greater than 2, respectively). A total of 43 out of 79 M^{pro} sequences (55%) showed nominally higher levels of mean constraint than WT (discounting significance), suggesting a lack of uniform selection pressure for active sites that are more or less constrained than WT (the fraction greater does not differ significantly from random deviation, $p = 0.16$, exact

https://dx.doi.org/10.1021/acs.biochem.0c00462
Biochemistry XXXX, XXX, XXX−XXX

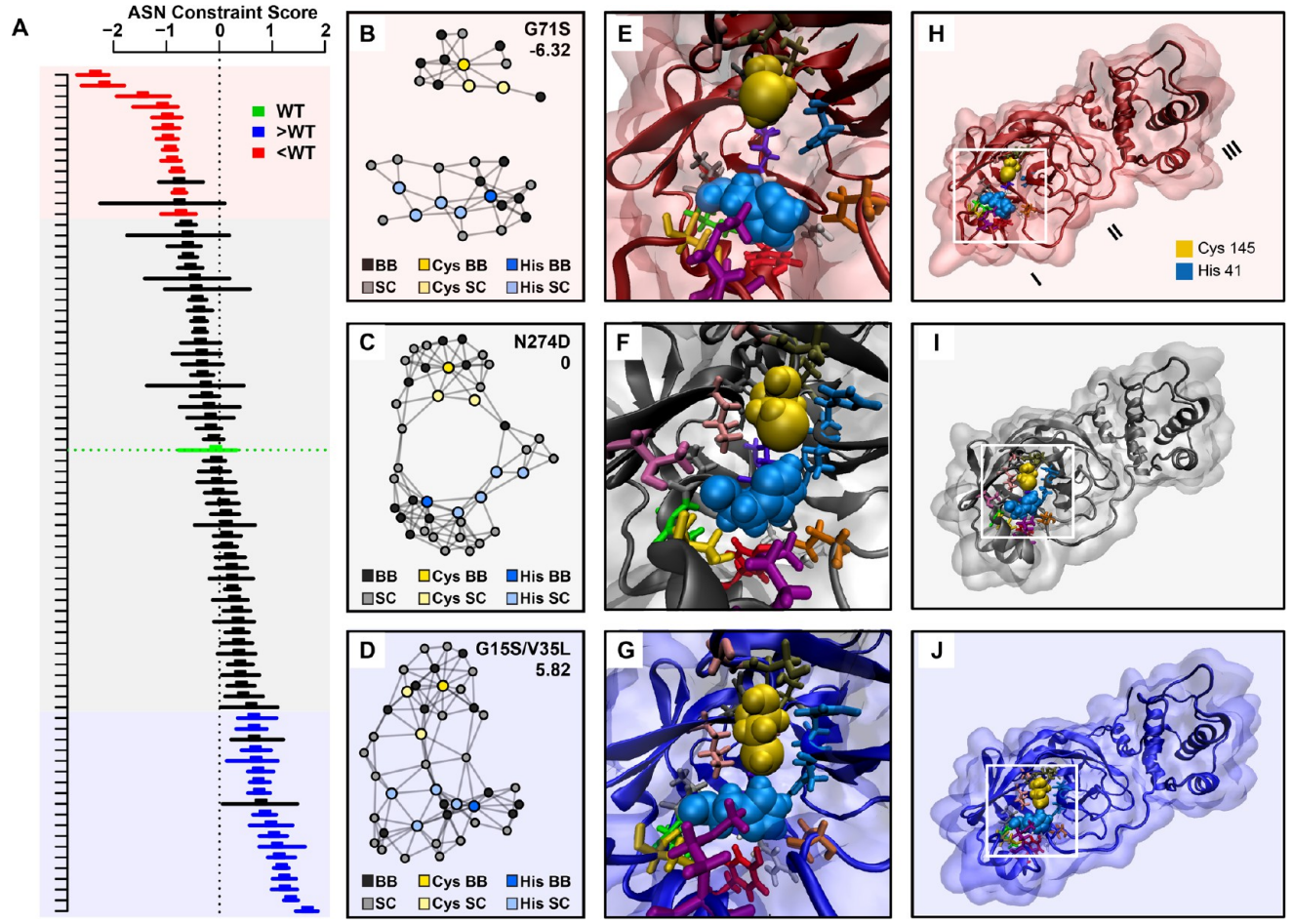


Figure 8. (A) Mean monomer active site constraint scores and 95% autocorrelation-corrected parametric bootstrap confidence intervals, by variant. Higher values indicate greater constraints on active site residues; red/blue intervals have t values versus WT (green) of at least ±2, indicating significant variation in average constraint across variants. (B) Minimum, (C) mean, and (D) maximum constraint ASNs over all frames. Low constraint conformations are characterized by no shared partners between the catalytic residues (colored nodes), while highly constrained conformations show cohesively reinforced contacts between them. (E, F, G) Active site residues and the surrounding residues making up the ASN for the proteins described in panels B, C, and D, respectively. (H, I, J) Full protein models for the same examples, with the active site regions indicated by white boxes.

binomial test). Thus, although we do not see evidence here of systematic selection for net changes in active site constraint in the monomeric state, we do see evidence that variants differ from each other and from WT in their average active site properties. Turning to the dimer trajectories, Figure 9A shows mean constraint scores for A and B chain active sites, with autocorrelation-corrected parametric bootstrap confidence intervals for each variant. Consistent with the hypothesis that dimerization can result in asymmetric modification of active site conformations,[20] we frequently observe large differences in the active site constraint between chains within the same variant; indeed, the root-mean-square difference in mean constraints within the variant is significantly larger than would be expected from the variation across sites alone (p = 0.048, permutation test), being approximately 1.65 times the standard deviation in mean constraint across variants. That said, such differentiation is not always present, with many trajectories (including WT) showing similar levels of constraint for both active sites, and it may not be necessary for function. (We observe that symmetric active site conformations were obtained in the WT dimer structure of Zhang et al.,[36] for instance.) Unlike the monomeric case, we see a greater trend here toward reduction in dimer active constraint scores versus WT, with 76% of sequences showing constraint levels lower than WT (p < 0.0001, exact binomial test). It is possible that this reflects a higher level of selection for looser active sites in the dimeric state per se. These differences should be considered when designing inhibitors that are tolerant of mutational change in M^{pro} over time. In particular, it is clear that the population of extant M^{pro} variants already possesses some phenotypic diversity in active site flexibility, potentially facilitating its ability to evolve around some types of inhibitors. The ability of M^{pro} dimers to sustain active sites with different levels of flexibility may also complicate inhibitor design and suggests the importance of inhibitors that are robust to variation in active site constraint.

Overall Variation in M^{pro} Active Site Conformations Is Related to Cys/His Contact. Given the diversity of active site conformations across monomeric/dimeric states, variants, and chains, it is useful to seek specific features that can be used to characterize those conformations. To examine this, we employ a kernel principal component analysis (kPCA) of a stratified sample of ASNs from the set of monomer and dimer trajectories, taking 25 evenly spaced frames from each trajectory (including both A and B ASNs in the dimeric case) for a total of 5925 conformations from the entire data

https://dx.doi.org/10.1021/acs.biochem.0c00462
Biochemistry XXXX, XXX, XXX−XXX

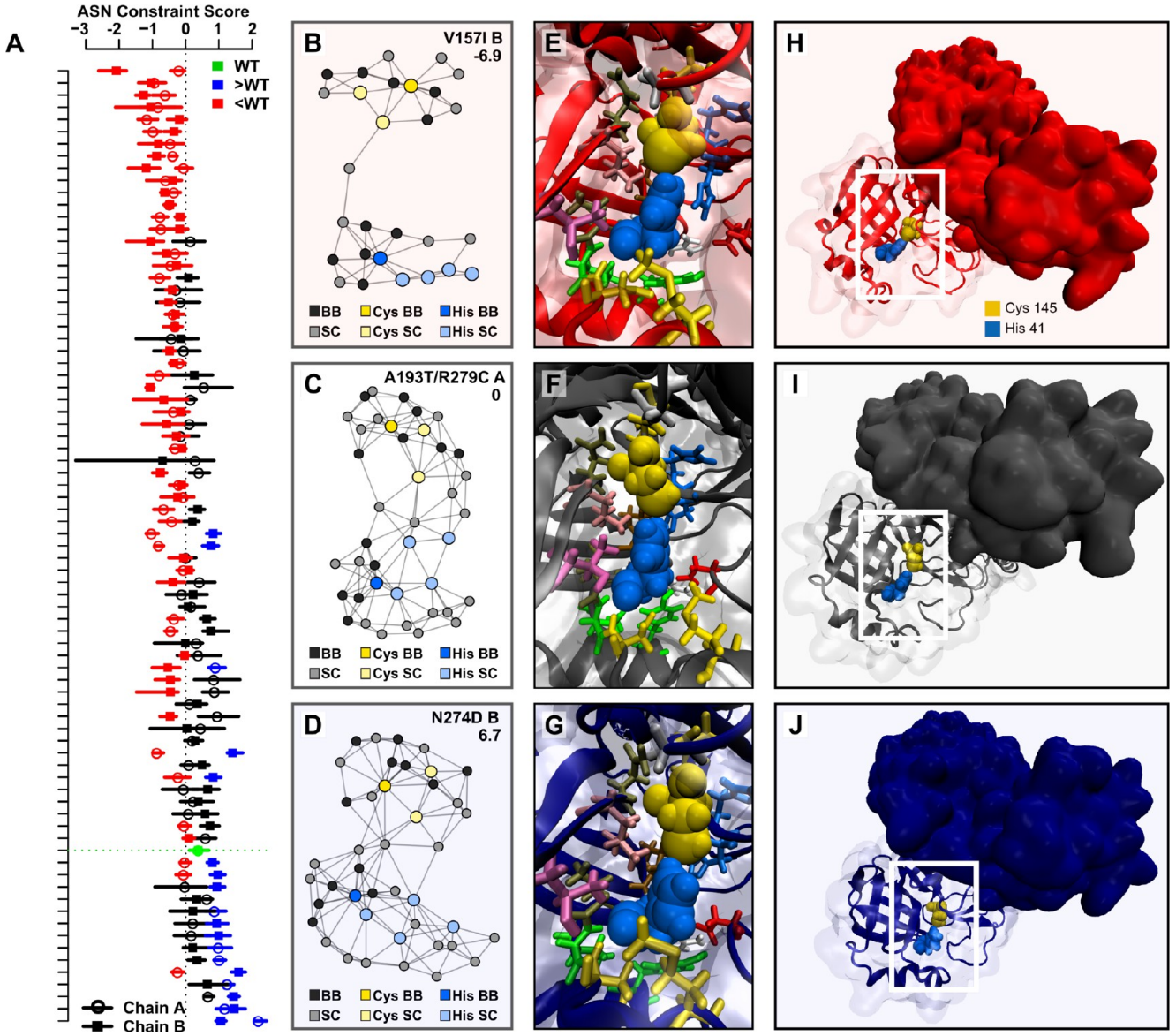


Figure 9. (A) Mean active site constraint scores and 95% autocorrelation-corrected parametric bootstrap confidence intervals for dimer A and B chain active sites, by variant. Red/blue intervals have t values versus WT (green) of at least ±2, indicating significant variation in average constraint across variants (comparisons made within the chain). (B) Minimum, (C) mean, and (D) maximum constraint ASNs over all structures. (E, F, G) Active site residues and the surrounding residues making up the ASN for the proteins described in panels B, C, and D, respectively. (H, I, J) Full protein models for the same examples, with the active site regions indicated by white boxes.

set. The feature space for our kernel is the set of all labeled subgraphs on the set of potential ASNs, allowing us to detect specific patterns of contacts among moieties that characterize subpopulations of conformations. As shown in Figure 10, the combined set of active site conformations falls into two fairly distinct clusters obliquely aligned with the first two principal components. These respective clusters correspond to a single feature, specifically the presence or absence of at least one contact between the catalytic cysteine and histidine residues; the cluster aligned with the first principal component consists of "open" conformations lacking such a contact, while the cluster aligned with the second principal component consists of "closed" conformations in which the contact is present.

Both clusters have an ellipsoidal form, with distance along the ellipsoids away from the origin corresponding to the average number of contacts per chemical group (mean degree). The cluster ellipsoids are also slightly oblique to one another, an orientation that arises from their joint respective correlation with the total number of moieties interacting with the active site residues (ASN size); this is shown in Figure 10 via the gradient of the size distribution, indicated as contour lines. These three properties (Cys/His contact, mean degree, ASN size) thus provide a parsimonious description of the most important axes of variation in active site structure and may be useful as targets for experimental investigation (e.g., via NMR). Interestingly, we observe very little association between conformation and monomer/dimer status (R^2 < 0.6% on the first PC, and no significant difference on the second-highest R^2 in the first 10 dimensions of 1.4%), indicating that the full range of active site conformations is observed in both monomer and dimer structures and at similar frequencies. Conformation is more strongly associated with the variant, the latter accounting for approximately 14% of the variance in the first PC and 7% in the second; adding interaction effects with

https://dx.doi.org/10.1021/acs.biochem.0c00462
Biochemistry XXXX, XXX, XXX−XXX

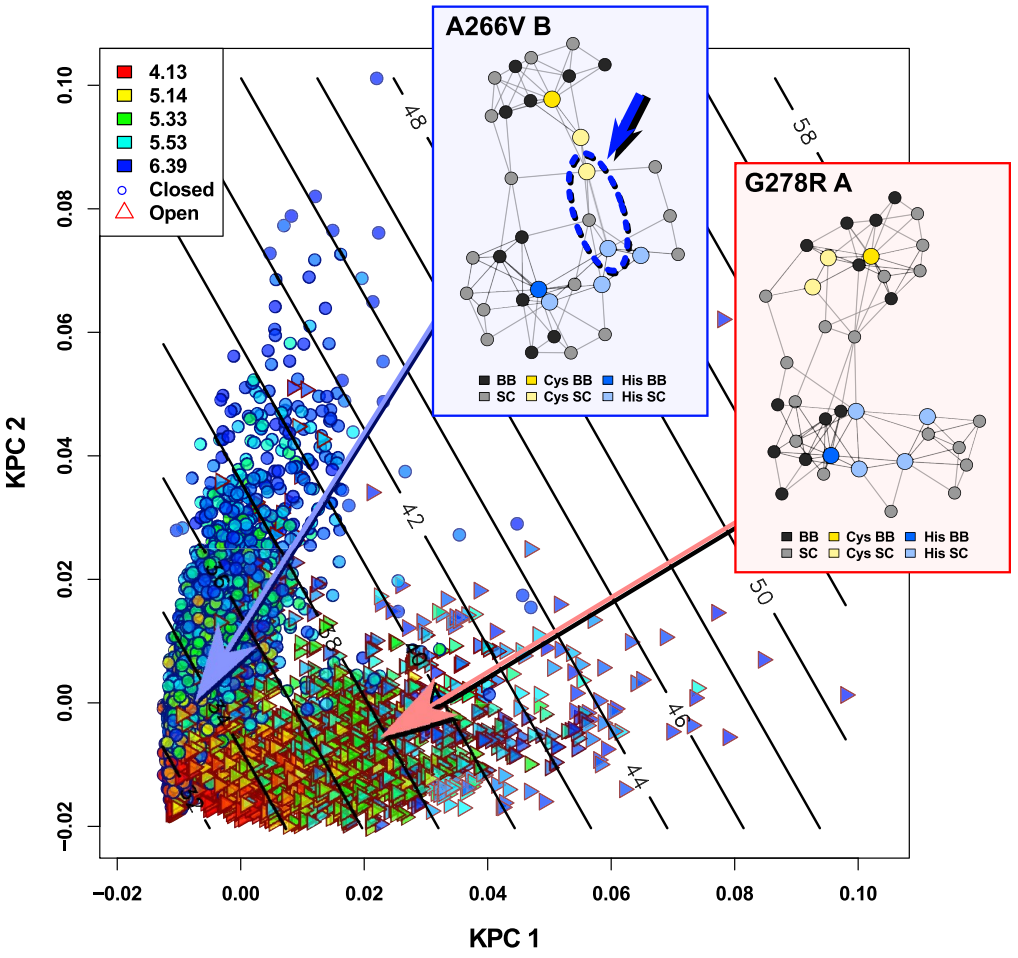


Figure 10. Kernel PCA solution for M^{pro} active site networks. Conformations fall into two clusters, corresponding to whether the Cys-His tie is present ("closed," circles) or absent ("open," triangles). The position within each cluster is strongly associated with mean degree (see point color) and secondarily with network size (gradient shown via contour lines). Inset networks illustrate centroids of each cluster; note the presence of the Cys/His interaction (highlighted) in the selected A226 V chain B conformation, which is absent in the selected G278R chain A conformation.

chain and monomer/dimer status increases this by 30% and 16%, respectively, suggesting that M^{pro} mutations tend to affect monomer and dimer structures differently. While the mutations observed to date do not radically alter active site conformation (this being a priori unlikely for functional mutations in any event), they do appear to exert a nontrivial influence on the distribution of conformational states.

CONCLUSION

For clinically relevant variants of M^{pro}, the observed variation so far is toward larger and more hydrophobic amino acids, leading to reduced structural cohesion on average in the variants relative to wild-type. Although mutations occur throughout the protein, the structural effects of those mutations appear to be far more localized, impacting primarily the protein's interdomain interfaces and several key loop regions near the substrate-binding pocket. These mutations appear to result in systematic variations in both global flexibility and in the extent to which the catalytic residues of the active site are constrained (a factor previously found to be related to substrate specificity in related systems). Our results suggest that M^{pro} may be currently subject to selection for enhanced global flexibility and that currently circulating M^{pro} variants represent a reservoir of phenotypic diversity in active site structure and dynamics that could facilitate an evolutionary response to certain classes of protease inhibitors. These findings are relevant to dimerization kinetics, substrate capture, and the development of resistance to inhibitors, as well as our understanding of SARS-CoV-2 more generally.

On a more methodological note, these results underscore the potential of comparative *in silico* studies to rapidly probe structural and functional consequences of genotypic variation in emerging diseases. Advances in both GPU-enabled hardware and molecular dynamics have made high-volume simulation studies feasible over short time horizons, giving us a powerful tool for selection of experimental targets. At the same time, comparative analysis of large volumes of trajectory data created by such simulation studies remains a challenge. Here, we have used both network analytic and machine learning techniques to identify potentially important sources of variation across trajectories. Although network analytic ideas have been used to study protein structures at least since 1993,[80] systematic use of combined network analytic and MD trajectories for comparative analysis of protein variants is more recent.[81−84] It is hoped that applications such as this one will inspire further development of this promising approach.

https://dx.doi.org/10.1021/acs.biochem.0c00462
Biochemistry XXXX, XXX, XXX−XXX

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.biochem.0c00462?goto=supporting-info.

Additional detail on the definitions of the PSNs used in this work; molecular models of representative variants; molecular models of WT and T225I monomers, showing interactions among relevant residues; molecular models of selected double variants; sites of K to R and R to C mutations, mapped on the wild-type dimer structure; sites of all known M^{pro} mutations, mapped on the wild-type structure; local variation index values for M^{pro} backbone torsion angle, by residue number; accession numbers, locations, and dates of collection of all M^{pro} variants referred to in this work; residue contacts for T/I225 in WT and T225I M^{pro}; mean cohesion scores (*k*-core number) and autocorrelation-corrected bootstrap standard errors for M^{pro} monomers by variant; mean cohesion scores (*k*-core number) for interfacial nodes and interfacial tie volumes with autocorrelation-corrected bootstrap standard errors by variant; mean active site network constraint scores by variant and chain, M^{pro} monomers; mean active site network constraint scores and autocorrelation-corrected bootstrap standard errors by variant and chain, M^{pro} dimers; MODELLER quality scores for all variants (monomer and dimer models) (PDF)

Uncompressed version of the tree depicted in Figure 1 (TXT)

Full acknowledgments for all sequences used in this work (PDF)

Accession Codes

SARS-CoV-2 main protease PDB ID: 6Y2E

■ AUTHOR INFORMATION

Corresponding Authors

Rachel W. Martin − *Department of Chemistry and Department of Molecular Biology and Biochemistry, University of California, Irvine, California 92697-2025, United States;* orcid.org/0000-0001-9996-7411; Email: rwmartin@uci.edu

Carter T. Butts − *California Institute for Telecommunications and Information Technology and Departments of Sociology, Statistics, Computer Science, and Electrical Engineering and Computer Science, University of California, Irvine, California 92697-3900, United States*; Email: buttsc@uci.edu

Authors

Thomas J. Cross − *Department of Chemistry, University of California, Irvine, California 92697-2025, United States*

Gemma R. Takahashi − *Department of Molecular Biology and Biochemistry, University of California, Irvine, California 92697-3900, United States*

Elizabeth M. Diessner − *Department of Chemistry and California Institute for Telecommunications and Information Technology, University of California, Irvine, California 92697-2025, United States*

Marquise G. Crosby − *Department of Molecular Biology and Biochemistry, University of California, Irvine, California 92697-3900, United States*

Vesta Farahmand − *Department of Chemistry, University of California, Irvine, California 92697-2025, United States*

Shannon Zhuang − *Department of Chemistry, University of California, Irvine, California 92697-2025, United States*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.biochem.0c00462

Funding

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

Notes

The authors declare no competing financial interest.

■ REFERENCES

(1) Wu, F., Zhao, S., Yu, B., Chen, Y.-M., Wang, W., Song, Z.-G., Hu, Y., Tao, Z.-W., Tian, J.-H., Pei, Y.-Y., Yuan, M.-L., Zhang, Y.-L., Dai, F.-H., Liu, Y., Wang, Q.-M., Zheng, J.-J., Xu, L., Holmes, E. C., and Zhang, Y.-Z. (2020) A new coronavirus associated with human respiratory disease in China. *Nature 579*, 265−269.

(2) Andersen, K. G., Rambaut, A., Lipkin, W. I., Holmes, E. C., and Garry, R. F. (2020) The proximal origin of SARS-CoV-2. *Nat. Med. 26*, 450−455.

(3) 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., and Shi, Z.-L. (2020) A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature 579*, 270−273.

(4) Liu, P., Chen, W., and Chen, J.-P. (2019) Viral metagenomics revealed sendai virus and coronavirus infection of Malayan pangolins (*Manis javanica*). *Viruses 11*, 979.

(5) Zhang, T., Wu, Q., and Zhang, Z. (2020) Probable pangolin origin of SARS-CoV-2 associated with the COVID-19 outbreak. *Curr. Biol. 30*, 1346−1351.

(6) Li, X., Giorgi, E. E., Marichannegowda, M. H., Foley, B., Xiao, C., Kong, X.-P., Chen, Y., Gnanakaran, S., Korber, B., and Gao, F. (2020) Emergence of SARS-CoV-2 through recombination and strong purifying selection. *Sci. Adv. 6*, No. eabb9153.

(7) Song, Z., Xu, Y., Bao, L., Zhang, L., Yu, P., Qu, Y., Zhu, H., Zhao, W., Han, Y., and Qin, C. (2019) From SARS to MERS, thrusting coronaviruses into the spotlight. *Viruses 11*, 59.

(8) Deeks, S. G., Smith, M., Holodniy, M., and Kahn, J. O. (1997) HIV-1 protease inhibitors: A review for clinicians. *JAMA, J. Am. Med. Assoc. 277*, 145−153.

(9) Sham, H. L., Kempf, D. J., Molla, A., Marsh, K. C., Kumar, G. N., Chen, C.-M., Kati, W., Stewart, K., Lal, R., Hsu, A., Betebenner, D., Korneyeva, M., Vasavanonda, S., McDonald, E., Saldivar, A., Wideburg, N., Chen, X., Niu, P., Park, C., Jayanti, V., Grabowski, B., Granneman, G. R., Sun, E., Japour, A. J., Leonard, J. M., Plattner, J. J., and Norbeck, D. W. (1998) ABT-378, a highly potent inhibitor of the human immunodeficiency virus protease. *Antimicrob. Agents Chemother. 42*, 3218−3224.

(10) Shuter, A. C. J. (2008) Lopinavir/ritonavir in the treatment of HIV-1 infection: a review. *Ther. Clin. Risk Manage. 4*, 1023−1033.

(11) Cao, B., Wang, Y., Wen, D., Liu, W., Wang, J., Fan, G., Ruan, L., Song, B., Cai, Y., Wei, M., Li, X., Xia, J., Chen, N., Xiang, J., Yu, T., Bai, T., Xie, X., Zhang, L., Li, C., Yuan, Y., Chen, H., Li, H., Huang, H., Tu, S., Gong, F., Liu, Y., Wei, Y., Dong, C., Zhou, F., Gu, X., Xu, J., Liu, Z., Zhang, Y., Li, H., Shang, L., Wang, K., Li, K., Zhou, X., Dong, X., Qu, Z., Lu, S., Hu, X., Ruan, S., Luo, S., Wu, J., Peng, L., Cheng, F., Pan, L., Zou, J., Jia, C., Wang, J., Liu, X., Wang, S., Wu, X., Ge, Q., He, J., Zhan, H., Qiu, F., Guo, L., Huang, C., Jaki, T., Hayden, F. G., Horby, P. W., Zhang, D., and Wang, C. (2020) A trial of lopinavir-

ritonavir in adults hospitalized with severe COVID-19. *N. Engl. J. Med. 382*, 1787.
(12) Snijder, E. J., Bredenbeek, P. J., Dobbe, J. C., Thiel, V., Ziebuhr, J., Poon, L. L. M., Guan, Y., Rozanov, M., Spaan, W. J. M., and Gorbalenya, A. E. (2003) Unique and conserved features of genome and proteome of SARS-coronavirus, an early split-off from the coronavirus group 2 lineage. *J. Mol. Biol. 331*, 991−1004.
(13) Shi, J., Wei, Z., and Song, J. (2004) Dissection study on the severe acute respiratory syndrome 3C-like protease reveals the critical role of the extra domain in dimerization of the enzyme: Defining the extra domain as a new target for design of highly specific protease inhibitors. *J. Biol. Chem. 279*, 24765−24773.
(14) Chou, C.-Y., Chang, H.-C., Hsu, W.-C., Lin, T.-Z., Lin, C.-H., and Chang, G.-G. (2004) Quaternary structure of the severe acute respiratory syndrome (SARS) coronavirus main protease. *Biochemistry 43*, 14958−14970.
(15) Fan, K., Wei, P., Feng, Q., Chen, S., Huang, C., Ma, L., Lai, B., Pei, J., Liu, Y., Chen, J., and Lai, L. (2004) Biosynthesis, purification, and substrate specificity of the severe acute respiratory syndrome coronavirus 3C-like proteinase. *J. Biol. Chem. 279*, 1637−1642.
(16) Lin, P. Y., Chou, C. Y., Chang, H. C., Hsu, W. C., and Chang, G. G. (2008) Correlation between dissociation and catalysis of SARS-CoV main protease. *Arch. Biochem. Biophys. 472*, 34−42.
(17) Zhong, N., Zhang, S., Zou, P., Chen, J., Kang, X., Li, Z., Liang, C., Jin, C., and Xia, B. (2008) Without its N-finger, the main protease of severe acute respiratory syndrome coronavirus can form a novel dimer through its C-terminal domain. *J. Virol. 82*, 4227−4234.
(18) 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., and Rao, Z. (2003) The crystal structures of severe acute respiratory syndrome virus main protease and its complex with an inhibitor. *Proc. Natl. Acad. Sci. U. S. A. 100*, 13190−13195.
(19) Chen, S., Chen, L., Tan, J., Chen, J., Du, L., Sun, T., Shen, J., Chen, K., Jiang, H., and Shen, X. (2005) Severe acute respiratory syndrome coronavirus 3C-like proteinase N-terminus is indispensable for proteolytic activity but not for enzyme dimerization: Biochemical and thermodynamic investigation in conjunction with molecular dynamics simulations. *J. Biol. Chem. 280*, 164−173.
(20) Chen, H., Wei, P., Huang, C., Tan, L., Liu, Y., and Lai, L. (2006) Only one protomer is active in the dimer of SARS 3C-like proteinase. *J. Biol. Chem. 281*, 13894−13898.
(21) Chen, S., Hu, T., Zhang, J., Chen, J., Chen, K., Ding, J., Jiang, H., and Shen, X. (2008) Mutation of Gly-11 on the dimer interface results in the complete crystallographic dimer dissociation of severe acute respiratory syndrome coronavirus 3C-like protease: Crystal structure with molecular dynamics simulations. *J. Biol. Chem. 283*, 554−564.
(22) Wright, E. S., Lakdawala, S. S., and Cooper, V. S. (2020) SARS-CoV-2 genome evolution exposes early human adaptations, *bioRxiv*.
(23) Jungreis, I., Sealfon, R., and Kellis, M. (2020) Sarbecovirus comparative genomics elucidates gene content of SARS-CoV-2 and functional impact of COVID-19 pandemic mutations, *bioRxiv* DOI: 10.1101/2020.06.02.130955.
(24) 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, M. D., Partridge, D. G., Evans, C. M., Freeman, T. M., de Silva, T. I., McDanal, C., Perez, L. G., Tang, H., Moon-Walker, A., Whelan, S. P., LaBranche, C. C., Saphire, E. O., Montefiori, D. C., Angyal, A., Brown, R. L., Carrilero, L., Green, L. R., Groves, D. C., Johnson, K. J., Keeley, A. J., Lindsey, B. B., Parsons, P. J., Raza, M., Rowland-Jones, S., Smith, N., Tucker, R. M., Wang, D., and Wyles, M. D. (2020) Tracking changes in SARS-CoV-2 spike: Evidence that D614G increases infectivity of the COVID-19 virus. *Cell 182*, 812.
(25) Bhattacharyya, C., Das, C., Ghosh, A., Singh, A. K., Mukherjee, S., Majumder, P. P., Basu, A., and Biswas, N. K. (2020) Global spread of SARS-CoV-2 subtype with spike protein mutation D614G is shaped by human genomic variations that regulate expression of TMPRSS2 and MX1 genes, *bioRxiv* DOI: 10.1101/2020.05.04.075911.
(26) Becerra-Flores, M., and Cardozo, T. (2020) SARS-CoV-2 viral spike G614 mutation exhibits higher case fatality rate. *Int. J. Clin. Pract. 74*, No. e13525.
(27) Portelli, S., Olshansky, M., Rodrigues, C. H., D'Souza, E. N., Myung, Y., Silk, M., Alavi, A., Pires, D. E., and Ascher, D. B. (2020) COVID-3D: An online resource to explore the structural distribution of genetic variation in SARS-CoV-2 and its implication on therapeutic development, *bioRxiv* DOI: 10.1101/2020.05.29.124610.
(28) Elbe, S., and Buckland-Merrett, G. (2017) Data, disease and diplomacy: GISAID's innovative contribution to global health. *Global Challenges 1*, 33−46.
(29) Van Rossum, G., and Drake, F. L., Jr (1995) *Python tutorial*, Vol. *620*, Centrum voor Wiskunde en Informatica Amsterdam.
(30) Cornish-Bowden, A. (1985) Nomenclature for incompletely specified bases in nucleic acid sequences: Recommendations 1984. *Nucleic Acids Res. 13*, 3021−3030.
(31) Inc., W. R. (2020) Mathematica, Champaign, IL. https://www.wolfram.com/mathematica.
(32) Edgar, R. C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res. 32*, 1792−1797.
(33) Kumar, S., Stecher, G., Li, M., Knyaz, C., and Tamura, K. (2018) MEGA X: Molecular Evolutionary Genetics Analysis across computing platforms. *Mol. Biol. Evol. 35*, 1547−1549.
(34) Saitou, N., and Nei, M. (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol. 4*, 406−425.
(35) Felsenstein, J. (1985) Confidence limits on phylogenies: An approach using the bootstrap. *Evolution 39*, 783−791.
(36) Zhang, L., Lin, D., Sun, X., Curth, U., Drosten, C., Sauerhering, L., Becker, S., Rox, K., and Hilgenfeld, R. (2020) Crystal structure of SARS-CoV-2 main protease provides a basis for design of improved α-ketoamide inhibitors. *Science 368*, 409−412.
(37) Webb, B., and Sali, A. (2016) Comparative protein structure modeling using Modeller. *Curr. Protoc. Bioinformatics 54*, 5.6.1−5.6.37.
(38) Olsson, M. H., Sondergaard, C. R., Rostkowski, M., and Jensen, J. H. (2011) PROPKA3: Consistent treatment of internal and surface residues in empirical pKa predictions. *J. Chem. Theory Comput. 7*, 525−537.
(39) Phillips, J. C., Braun, R., Wang, W., Gumbart, J., Tajkhorshid, E., Villa, E., Chipot, C., Skeel, R. D., Kalé, L., and Schulten, K. (2005) Scalable molecular dynamics with NAMD. *J. Comput. Chem. 26*, 1781−1802.
(40) Best, R. B., Zhu, X., Shim, J., Lopes, P. E. M., Mittal, J., Feig, M., and MacKerell, A. D. (2012) Optimization of the additive CHARMM all-atom protein force field targeting improved sampling of the backbone ϕ, ψ and side-chain $\chi(1)$ and $\chi(2)$ dihedral angles. *J. Chem. Theory Comput. 8*, 3257−3273.
(41) Jorgensen, W. L., Chandrasekhar, J., Madura, J. D., Impey, R. W., and Klein, M. L. (1983) Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys. 79*, 926−935.
(42) Martyna, G. J., Tobias, D. J., and Klein, M. L. (1994) Constant pressure molecular dynamics algorithms. *J. Chem. Phys. 101*, 4177−4189.
(43) Feller, S. E., Zhang, Y., Pastor, R. W., and Brooks, B. R. (1995) Constant pressure molecular dynamics simulation: The Langevin piston method. *J. Chem. Phys. 103*, 4613−4621.
(44) Handcock, M. S., Hunter, D. R., Butts, C. T., Goodreau, S. M., and Morris, M. (2008) statnet: Software Tools for the Representation, Visualization, Analysis and Simulation of Network Data. *J. Stat. Softw. 24*, 1−11.
(45) Butts, C. T. (2008) network: a Package for Managing Relational Data in R. *J. Stat. Softw. 24*, 1.
(46) Butts, C. T. (2008) Social Network Analysis with sna. *J. Stat. Softw. 24*, 1.

https://dx.doi.org/10.1021/acs.biochem.0c00462
Biochemistry XXXX, XXX, XXX−XXX

(47) Idé, J. (2017) *Rpdb: Read, Write, Visualize and Manipulate PDB Files*, R Package Version 2.3.
(48) Grant, B. J., Rodrigues, A. P. C., ElSawy, K. M., McCammon, J. A., and Caves, L. S. D. (2006) Bio3D: An R package for the comparative analysis of protein structures. *Bioinformatics 22*, 2695−2696.
(49) R Core Team (2020) *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.
(50) Benson, N. C., and Daggett, V. (2012) A chemical group graph representation for efficient high-throughput analysis of atomistic protein simulations. *J. Bioinf. Comput. Biol. 10*, 1250008.
(51) Alvarez, S. (2013) A Cartography of the van der Waals Territory. *Dalton Trans. 42*, 8617−8636.
(52) Duong, V. T., Unhelkar, M. H., Kelly, J. E., Kim, S., Butts, C. T., and Martin, R. W. (2018) Network analysis provides insight into active site flexibility in esterase/lipases from the carnivorous plant *Drosera capensis*. *Integr. Biol. 10*, 768−779.
(53) Seidman, S. B. (1983) Network Structure and Minimum Degree. *Social Networks 5*, 269−287.
(54) Wasserman, S., and Faust, K. (1994) *Social network analysis: methods and applications*, Vol. *8*, Cambridge University Press.
(55) Unhelkar, M. H., Duong, V. T., Enendu, K. N., Kelly, J. E., Tahir, S., Butts, C. T., and Martin, R. (2017) Structure prediction and network analysis of chitinases from the Cape sundew, *Drosera capensis*. *Biochim. Biophys. Acta, Gen. Subj. 1861*, 636−643.
(56) Efron, B., and Tibshirani, R. (1993) *An Introduction to the Bootstrap*, Chapman and Hall, London.
(57) Scholkopf, B., and Smola, A. J. (2001) *Learning with Kernels: Support Vector Machines Regularization, Optimization, and Beyond*, MIT Press, Cambridge, MA.
(58) Kowalczyk, A., Smola, A. J., and Williamson, R. C. (2001) Kernel Machines and Boolean Functions. In *Proceedings of the 14th International Conference on Neural Information Processing Systems: Natural and Synthetic*, pp 439−446, Cambridge, MA.
(59) Kwok, J. T., and Tsang, I. W. (2003) The Pre-Image Problem in Kernel Methods. In *Proceedings of the Twentieth International Conference on Machine Learning (ICML-2003)*, Washington DC.
(60) Eddelbuettel, D., and Balamuta, J. J. (2017) Extending R with C ++: A Brief Introduction to Rcpp. *PeerJ. Preprints 5*, No. e3188v1.
(61) Wang, C., Liu, Z., Chen, Z., Huang, X., Xu, M., He, T., and Zhang, Z. (2020) The establishment of reference sequence for SARS-CoV-2 and variation analysis. *J. Med. Virol. 92*, 667−674.
(62) Bogatyreva, N. S., Finkelstein, A. V., and Galzitskaya, O. V. (2006) Trend of amino acid composition of proteins of different taxa. *J. Bioinf. Comput. Biol. 4*, 597−608.
(63) Hadfield, J., Megill, C., Bell, S. M., Huddleston, J., Potter, B., Callender, C., Sagulenko, P., Bedford, T., and Neher, R. A. (2018) Nextstrain: real-time tracking of pathogen evolution. *Bioinformatics 34*, 4121−4123.
(64) Wu, F., Zhao, S., Yu, B., Chen, Y.-M., Wang, W., Song, Z.-G., Hu, Y., Tao, Z.-W., Tian, J.-H., Pei, Y.-Y., Yuan, M.-L., Zhang, Y.-L., Dai, F.-H., Liu, Y., Wang, Q.-M., Zheng, J.-J., Xu, L., Holmes, E. C., and Zhang, Y.-Z. (2020) A new coronavirus associated with human respiratory disease in China. *Nature 579*, 265−269.
(65) Xiao, K., Zhai, J., Feng, Y., Zhou, N., Zhang, X., Zou, J.-J., Li, N., Guo, Y., Li, X., Shen, X., Zhang, Z., Shu, F., Huang, W., Li, Y., Zhang, Z., Chen, R.-A., Wu, Y.-J., Peng, S.-M., Huang, M., Xie, W.-J., Cai, Q.-H., Hou, F.-H., Chen, W., Xiao, L., and Shen, Y. (2020) Isolation and characterization of 2019-nCoV-like coronavirus from Malayan pangolins. *Nature 583*, 286.
(66) Roser, M., Ritchie, H., Ortiz-Ospina, E., and Hasell, J. (2020) Coronavirus Disease (COVID-19). *Our World in Data*.
(67) Kyte, J., and Doolittle, R. F. (1982) A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol. 157*, 105−132.
(68) Brüne, D., Andrade-Navarro, M. A., and Mier, P. (2018) Proteome-wide comparison between the amino acid composition of domains and linkers. *BMC Res. Notes 11*, 117.
(69) Echols, N., Harrison, P., Balasubramanian, S., Luscombe, N. M., Bertone, P., Zhang, Z., and Gerstein, M. (2002) Comprehensive analysis of amino acid and nucleotide composition in eukaryotic genomes, comparing genes and pseudogenes. *Nucleic Acids Res. 30*, 2515−2523.
(70) Shimoni, L., and Glusker, J. (1995) Hydrogen bonding motifs of protein side chains: Descriptions of binding of arginine and amide groups. *Protein Sci. 4*, 65−74.
(71) Gallivan, J., and Dougherty, D. (1999) Cation-π interactions in structural biology. *Proc. Natl. Acad. Sci. U. S. A. 96*, 9459−9464.
(72) Crowley, P., and Golovin, A. (2005) Cation-π interactions in protein-protein interfaces. *Proteins: Struct., Funct., Genet. 59*, 231−239.
(73) Vondvorášek, J., Mason, P., Heyda, J., Collins, K., and Jungwirth, P. (2009) The molecular origin of like-charge arginine-arginine pairing in water. *J. Phys. Chem. B 113*, 9041−9045.
(74) Hsu, M.-F., Kuo, C.-J., Chang, K.-T., Chang, H.-C., Chou, C.-C., Ko, T.-P., Shr, H.-L., Chang, G.-G., Wang, A. H.-J., and Liang, P.-H. (2005) Mechanism of the maturation process of SARS-CoV 3CL protease. *J. Biol. Chem. 280*, 31257−31266.
(75) Ménard, R., and Storer, A. C. (1992) Oxyanion hole interactions in serine and cysteine proteases. *Biol. Chem. Hoppe-Seyler 373*, 393−400.
(76) Shi, J., Sivaraman, J., and Song, J. (2008) Mechanism for controlling the dimer-monomer switch and coupling dimerization to catalysis of the severe acute respiratory syndrome coronavirus 3C-like protease. *J. Virol. 82*, 4620−4629.
(77) Cross, T. J., Takahashi, G. R., Diessner, E. M., Crosby, M. G., Farahmand, V., Zhuang, S., Butts, C. T., and Martin, R. W. (2020) Sequence characterization and molecular modeling of clinically relevant variants of the SARS-CoV-2 main protease. In *bioRxiv* DOI: 10.1101/2020.05.15.097493
(78) Amamuddy, O. S., Verkhivker, G. M., and Bishop, Ö. T. (2020) Impact of emerging mutations on the dynamic properties the SARS-CoV-2 main protease: An in silico investigation. In *bioRxiv* DOI: 10.1101/2020.05.29.123190.
(79) Sheik Amamuddy, O., Verkhivker, G. M., and Tastan Bishop, O. (2020) Impact of early pandemic stage mutations on molecular dynamics of SARS-CoV-2 Mpro. In *J. Chem. Inf. Model*, in press DOI: 10.1021/acs.jcim.0c00634.
(80) Aszói, A., and Taylor, W. (1993) Connection topology of proteins. *Bioinformatics 9*, 523−529.
(81) James, K. A., and Verkhivker, G. M. (2014) Structure-based network analysis of activation mechanisms in the ErbB family of receptor yyrosine kinases: The regulatory spine residues are global mediators of structural stability and allosteric interactions. *PLoS One 9*, No. e113488.
(82) Bhakat, S., Martin, A. J. M., and Soliman, M. E. S. (2014) An integrated molecular dynamics, principal component analysis and residue interaction network approach reveals the impact of M184V mutation on HIV reverse transcriptase resistance to lamivudine. *Mol. BioSyst. 10*, 2215−2228.
(83) Brown, D. K., and Tastan Bishop, O. (2019) Role of structural bioinformatics in drug discovery by computational SNP analysis: Analyzing variation at the protein level. *Glob. Heart 12*, 151−161.
(84) Grazioli, G., Martin, R. W., and Butts, C. T. (2019) Comparative exploratory analysis of intrinsically disordered protein dynamics using machine learning and metwork analytic methods. *Front. Mol. Biosci. 6*, 1.

https://dx.doi.org/10.1021/acs.biochem.0c00462
Biochemistry XXXX, XXX, XXX−XXX