Title
Elucidation of Cryptic and Allosteric Pockets within the SARS-CoV-2 Main Protease.

Permalink
https://escholarship.org/uc/item/21d752gk

Authors
Sztain, Terra
Amaro, Rommie
McCammon, J Andrew

Publication Date
2021-05-03

DOI
10.1021/acs.jcim.1c00140

Peer reviewed
Elucidation of Cryptic and Allosteric Pockets within the SARS-CoV-2 Main Protease

Terra Sztain,* Rommie Amaro, and J. Andrew McCammon

ABSTRACT: The SARS-CoV-2 pandemic has rapidly spread across the globe, posing an urgent health concern. Many quests to computationally identify treatments against the virus rely on in silico small molecule docking to experimentally determined structures of viral proteins. One limit to these approaches is that protein dynamics are often unaccounted for, leading to overlooking transient, druggable conformational states. Using Gaussian accelerated molecular dynamics to enhance sampling of conformational space, we identified cryptic pockets within the SARS-CoV-2 main protease, including some within regions far from the active site. These simulations sampled comparable dynamics and pocket volumes to conventional brute force simulations carried out on two orders of magnitude greater timescales.

INTRODUCTION

In December 2019, the World Health Organization learned of a novel coronavirus that has since rapidly spread, leading to a global pandemic. The SARS-CoV-2 virus, causing the disease named COVID-19, has exceeded 100 million cases worldwide, taking over 2 million lives as of January 30, 2021.1 The genetic and structural similarity of the SARS-CoV-2 virus to the agents of the severe acute respiratory syndrome coronavirus (SARS-CoV) epidemic in 2003 and Middle East respiratory syndrome coronavirus (MERS-CoV) epidemic in 2012 provide a basis of information for understanding and ultimately treating or preventing this disease; however, no highly effective antiviral drug exists for any human-infecting coronavirus.

Proteases are responsible for activating viral proteins for particle assembly and have proved successful targets for antiviral agents; most notable are the protease inhibitors used to treat HIV and Hepatitis C.2−5 The main protease of SARS-CoV-2, called $\text{M}^{\text{pro}}$, or 3CL$^{\text{pro}}$ encoded by the nsp5 gene,4 was the first SARS-CoV-2 protein deposited to the protein databank (PDB) on January 26th 2020.5 This structure was crystallized with a covalent inhibitor (N3) identified from computer-aided drug design and validated biochemically. Currently, over one hundred $\text{M}^{\text{pro}}$ structures exist in the PDB and massive efforts to discover a successful inhibitor are underway.6−8

$\text{M}^{\text{pro}}$ is highly similar to 3CL proteases from other coronaviruses in sequence, structure, and function.9,10 First, $\text{M}^{\text{pro}}$ autocatalytically cleaves itself from the SARS-CoV-2 polypeptide and then forms a homodimer to subsequently cleave at 11 distinct sites, while the papain-like protease, PL$^{\text{pro}}$, cleaves at three sites, allowing the viral proteins to fold and perform their functions.11−13 Proteolytic cleavage is catalyzed by residues H41 and C145 of the catalytic cysteine dyad. These active site residues are located at the N-terminal globular domain of $\text{M}^{\text{pro}}$, which contains an anti-parallel beta barrel reminiscent of trypsin-like serine proteases, while the C-terminal domain consists of five alpha helices.13 Upon dimerization, $\text{M}^{\text{pro}}$ resembles the shape of a heart symbol (Figure 1A).

Received: February 7, 2021

This article is made available via the ACS COVID-19 subset for unrestricted RESEARCH re-use and analyses in any form or by any means with acknowledgement of the original source. These permissions are granted for the duration of the World Health Organization (WHO) declaration of COVID-19 as a global pandemic.
Figure 1. Pocket analysis of GaMD simulations of MPro. (A) Active site, dimer interface, and distal site definitions for pocket calculations. The inclusion sphere was centered on the center of mass (COM) of the top and bottom protein domains for active and distal sites with 12 and 10 Å radii, respectively. A 10 Å-radius sphere was used for the dimer interface calculation. (B–D) Histograms showing the distribution of pocket volumes calculated from the aggregate 6 μs GaMD simulations.

Figure 2. Cryptic pockets identified from GaMD simulations. (A–C) Overlay of several frames highlighting loop dynamics leading to various pocket conformations for each site. (D–F) Overlay of several pockets shown outlined in mesh, colored in varying shades of pink for the active site, blue for the distal site, and orange for the interface. (G–I) Computational solvent mapping of “hot spots” to pockets at each site.
Covalently bound N3 sampled greater mean volumes for the active site and dimer interface than non-covalent, which sampled greater mean volumes for the distal site (Figure S3). In a few simulations, the non-covalent ligand escaped the active site pocket (Figure S4). In all simulations, the ligand rearranges compared to the crystal structure. The non-covalent ligands quickly decrease in total number of native contacts and increase in non-native contacts. The covalent ligands slightly decrease in native contacts but have a similar increase in non-native contacts. (Figures S5 and S6). These simulations revealed significant loop dynamics (Figure 2A–C) leading to the distinct pocket conformations (Figure 2D–F and Movie S1) with varying active site water occupancy (Figure S7).

Several metrics were used to assess the properties of these pockets. The PockDrug\textsuperscript{19} webserver predicts druggability probability based on a variety of factors including geometry, hydrophobicity, and aromaticity, among others. The FTMap\textsuperscript{20} computational solvent mapping webserver helps determine “hot spots” and can aid in fragment-based drug design. For each site, the sampled pocket volumes were clustered into 10 populations, and a frame corresponding to that volume was analyzed. The most populated volumes for the active site, dimer interface, and distal site were 114, 220, and 10 Å\textsuperscript{3}, respectively. The 10 active site pockets gave predicted druggabilities between 0.18 and 0.86 and solvents mapped to seven of the pockets (Figure S8). The dimer interface pockets gave higher druggability prediction values, between 0.5 and 0.93, with solvent mapping to all 10 structures (Figure S9). The distal site pockets gave the highest predicted druggability, between 0.37 and 1.00 with six out of ten pockets giving predicted druggability values greater than 0.90. Solvents mapped to seven out of the ten pockets (Figure S10). Notably, none of the solvent mapping to any of the sites showed overlapping results, indicating that each structure may provide distinct opportunities for designing inhibitors. The pocket volume, predicted druggability, and solvent mapping were not directly correlated, as expected, since each calculation considers different metrics. Overall, these analyses indicate that the simulations reveal a diverse set of conformations that are likely to serve as viable targets for in silico drug development.

Dimer association is necessary for catalytic activity of the protease;\textsuperscript{5} therefore, it is reasonable to assume that binding of a molecule, which prevents this association, would inactivate
the protease. The significance of the distal regions is less well understood, though computational investigations support a potential allosteric role.\textsuperscript{21,22} To study this further, we examined the correlated motions of each residue throughout the simulations. In the monomer simulation, the active site region was positively correlated to the distal site region and negatively correlated to the region that connects the two, whereas the inverse was observed in the dimer simulation (Figure 3 and Figures S11 and S12). These correlations resulted in primary motions, detected by principal component analysis, comprising hinge-like in the monomer versus twisting in the dimer (Figure S13). These motions were similar in both the apo and N3 bound simulations. In the dimer simulations, the active site of chain A was slightly correlated to the active site of chain B, and the distal regions were highly anticorrelated, twisting in opposite directions. (Figure S13). Regardless of positive or negative correlation values, the distal and active sites show a dynamic interdependence on each other in both the apo and dimer forms.

\textbf{GaMD Enhancement of M\textsuperscript{pro} Conformational Space.}\n
Previous benchmarks have demonstrated that GaMD can enhance sampling by multiple orders of magnitude.\textsuperscript{23} To test the sampling enhancement compared to conventional MD simulations, we compared the 1 \(\mu\)s aggregate GaMD simulation from five independent 200 \(\mu\)s simulations carried out at Riken\textsuperscript{24} and 100 \(\mu\)s simulations D.E. Shaw Research (DESRES).\textsuperscript{25} We additionally performed and compared those to 1 \(\mu\)s of continuous GaMD simulation and compared 200 ns cMD to 200 ns GaMD. The ranges of pocket volumes for each site were highly similar, as detailed in Figure S14. Even the short conventional simulations sampled comparable pocket volumes, indicating that this metric may not be the most effective measure of conformational sampling. Principal component analysis (PCA) is frequently used as a metric for conformational diversity as it decreases the dimensionality to maximize the dataset variance.\textsuperscript{26,27} PCA was indeed more consistent with expectations where increasing simulation time corresponded to larger regions of space covered by plotting the first two principal components (Figure 4).

While the GaMD simulations run continuously for 1 \(\mu\)s only covered 54\% of the conformational space covered by the 100 \(\mu\)s DESRES, the 1 \(\mu\)s aggregate GaMD from five simulations run in parallel covered 80\% of the DESRES space. The 10 \(\mu\)s Riken simulations covered 94\% of that space (Figure 4). It is worth noting that each of the simulations with the most extensive sampling (DESRES, Riken, GaMD 5 \(\times\) 200 ns) reached areas in the PC1-PC2 subspace that neither of the other two reached. This just confirms that none of these simulations achieve comprehensive sampling, which, in the ultimate case, would include extensively unfolded conformations. Each of these simulations reveals cryptic binding sites not seen in shorter, conventional MD simulations, let alone in the experimental structures, and some sites are seen in one but not the other simulations. Though perhaps not intuitive at first glance, it is reasonable that many simulations in parallel would have a greater chance of sampling diverse space than one long simulation, which may settle into a favorable low energy conformation, even given the boost added by the GaMD, which is restricted by a

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{figure4.png}
\caption{Conformational sampling of GaMD and conventional MD. Each plot shows the first and second principal components, projected from a mass weighted matrix of backbone atom positions of the apo dimer, colored by probability density. (A) 200 ns of GaMD simulation. (B) 1 \(\mu\)s of GaMD aggregated from five independent 200 \(\mu\)s simulation. (C) 1 \(\mu\)s of continuous GaMD simulation. (D) 200 ns of cMD. (E) 10 \(\mu\)s of cMD performed by Riken. (F) 100 \(\mu\)s of cMD performed by D.E. Shaw Research (DESRES). The area of a 2D histogram of the PC1 vs PC2 datapoints from each simulation is labeled as a percentage relative to the 100 \(\mu\)s DE. Shaw Research (DESRES) conventional MD simulation.}
\end{figure}
threshold. Representative structures from dense populations of the PCA plots were further analyzed structurally and through computational solvent mapping (Figure S15). Additionally, RMSD and RMSF comparison revealed that the majority of the conformational deviations occur within the helices and loops lining the active site pocket (Figures S16–S17, and S21–S23), which was also visualized in the representative structures from populations determined from PCA (Figure S15). Taken together, the range of pocket volumes, structural conformations, and dynamics were similar between short GaMD simulations and brute force conventional simulations. The plethora of available structures from all of these studies is expected to be useful for in silico inhibitor development for targeting Mpro.

## CONCLUSIONS

GaMD simulations of the SARS-CoV-2 main protease, Mpro, revealed cryptic pockets not detectable from the crystal structure. In addition to characterizing the vast conformational landscape of the active site and dimer interface, a distal region was explored as a potential allosteric site. GaMD sampled conformational space more efficiently than brute force simulations, with 1 μs simulation data providing comparable results to 10 μs simulations carried out by Riken and 100 μs simulations carried out by DE Shaw Research, offering a more widely accessible simulation strategy. These structures, along with the millisecond simulations from Folding@home, can serve as a basis for in silico docking to identify Mpro inhibitors for COVID-19 treatment.

## METHODS

### Simulation Preparation

The first crystal structure of SARS-CoV-2 main protease, PDB ID 6LU7, was used as a starting point for all simulations. Protonation states of titratable residues were determined using the H++ webserver with 0.15 M salinity, 10 internal dielectric constant, 80 external dielectric constant, and pH 7.4. Histidines 64, 80, and 164 were thus named HID to indicate delta nitrogen protonation before protonation of entire protein using tleap from AmberTools 18. Six systems were simulated as shown in Figure S1. For apo simulations, the inhibitor was deleted from the PDB file. For dimer simulations, the PyMOL symexp command was used to generate initial coordinates for the second chain based on symmetry. The ff14SB forcefield was used for proteogenic residues. For simulations with the N3 bound, the antechamber package of AmberTools 18 was used to generate GAFF forcefield parameters for N3. Gaussian 09 was used to calculate partial charges of atoms according to the RESP method with the HF/6-31G* level of theory. For covalent simulations, the cysteine 145 was treated as an unnatural amino acid and capped with N-methyl and C-acetyl groups for charge calculations. A TIP3P isometric water box was added with at least 10 Å buffer between solute and edges of the box. Enough Na+ was added to neutralize the system, and then Na+ and Cl– ions were added to a final concentration of 150 mM. Minimization was carried out in two steps. First, the solute was restrained with a 500 kcal/mol restraint force to minimize the solvent for 10,000 cycles followed by unrestrained minimization for 300,000 cycles. Next, the system was heated to 310 K over 350 ps using an isothermal-isobaric (NPT) ensemble followed by isothermal-isobaric (NPT) equilibration for 1 ns.

### Gaussian Accelerated Molecular Dynamics

Gaussian accelerated molecular dynamics (GaMD) pmemd.cuda implementation of Amber 18 was used to generate five independent trajectories of 200 ns, an aggregate of 1 μs for each system. The dual boost method was employed, adding a biasing force to both the total and dihedral potential energy. The threshold energy was set to the upper bound. GaMD production of equilibrated systems was carried out in five stages. First, 200 ps of preparatory conventional MD simulation was carried out, without any statistics collected. Second, 1 ns of conventional MD was carried out to collect potential statistics \( V_{\text{max}} \), \( V_{\text{avg}} \), \( V_{\text{std}} \), and \( \sigma V \). Next, 800 ps of GaMD was carried out with a boost potential applied with fixed parameters. Then, 50 ns of GaMD was carried out with updated boost parameters, and finally, 150 ns of GaMD was carried out with fixed boost parameters. All production simulations were carried out using NVT, with periodic boundary conditions and 2 fs timesteps. The SHAKE algorithm was used to restrain nonpolar hydrogen bonds and TIP3P water molecules. The Particle Mesh Ewald method was used for electrostatic interactions with a 1 Å cutoff for nonbonded interactions. A Langevin thermostat was used for temperature regulation with a collision frequency of 5 ps–1.

### Pocket Analysis

For each system, the aggregated simulation was clustered based on the RMSD from the first frame using the cpptraj hierarchical algorithm to obtain 100 distinct structures for pocket calculation. Three regions were defined for individual calculations: the active site pocket, distal site pocket, and dimer interface. Pocket volume was calculated using POVME 2 with a 1 Å grid spacing. The point inclusion method was used for electrostatic interactions with a 1 Å cutoff for nonbonded interactions. A Langevin thermostat was used for temperature regulation with a collision frequency of 5 ps–1. The density of oxygen centers map was used to visualize results in VMD.

To quantify druggability, we have used the PockDrug webserver, which predicts druggability probability based on a variety of factors including geometry, hydrophobicity, and aromaticity, among others. Additionally, we have used the FTMap computational solvent mapping webserver to determine druggable “hot spots.” Pocket volumes calculated by POVME from each of the three regions were clustered into 10 populations, and representative frames for each were selected for analysis. The pocket estimation method was used for PockDrug, and the druggability probability of the pocket
that aligned best to the POVME calculation sphere was reported.

**Principal Component Analysis.** Principal motions were calculated using cpptraj, as described previously, from projection of displacement vectors of each of the backbone atoms onto a diagonalized mass-weighted covariance matrix after rms fitting of every atom except protons to the first frame. Residue correlation matrices were also generated this way. For the graphical abstract, reweighting of the potential energy was carried out on one 200 ns simulation of the one N3 non-covalently bound dimer using the PyReweighting toolkit. Reweighting was carried out using the Maclaurin expansion series expansion to the 10th order with a bin size of 6 and maximum energy of 100 based on the first two principal components. One conventional 200 ns MD simulation was performed for comparison and was outlined in the figure. For the conformational sampling comparison shown in Figure 4, the dimer–apo system was used for consistency with Riken and DESRES simulations. The colors of the histograms are normalized so the same number of datapoints (frames) is considered. Therefore, 40,000 frames were used for all of the in house simulations, and 50,000 frames were used for the Riken and DESRES simulations. The color bars were set to maximum values of 0.0008 and 0.0010, respectively. This results in densely localized (yellow color) in the short simulations versus less density (blue color) and more spread out in the longer and enhanced simulations. Visualization of the first normal mode was carried out using the Normal Mode Wizard plugin of VMD.

**DATA AND SOFTWARE AVAILABILITY**

In accordance with the community principles around open sharing of COVID19 simulation data, all simulation input files and data are available at https://covid19.molssi.org through the NSF MolSSI COVID19 Molecular Structure and Therapeutics Hub.

**ASSOCIATED CONTENT**

**Supporting Information**

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jcim.1c00140. Values for RMSD of 275 of 3M protein crystal structures compared to PDB 6LU7 (XLSX)

Active site pocket dynamics sampled with GaMD (MP4)

Supplementary Figures S1–S26; legend for Supplementary Movie S1, and legend for Supplementary Data S1 (PDF)

**AUTHOR INFORMATION**

**Corresponding Author**

Terra Sztain — Department of Chemistry and Biochemistry, University of California, San Diego, La Jolla, California 92093, United States; orcid.org/0000-0002-1327-8541; Email: tsztainp@ucsd.edu

**Authors**

Rommie Amaro — Department of Chemistry and Biochemistry, University of California, San Diego, La Jolla, California 92093, United States; orcid.org/0000-0002-9275-9553

J. Andrew McCammon — Department of Chemistry and Biochemistry and Department of Pharmacology, University of California, San Diego, La Jolla, California 92093, United States; orcid.org/0000-0003-3065-1456

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jcim.1c00140

**Notes**

The authors declare no competing financial interest.

**ACKNOWLEDGMENTS**

We thank Prof. Michael D. Burkart for advisement, and the NSF GRFP for supporting T.S. under grant number DGE-1650112. Simulations were carried out using GPU nodes of the Triton Shared Computing Cluster (TSCC) at the San Diego Supercomputer Center (SDSC). This work was supported by NIH GM031749.

**REFERENCES**

(1) COVID-19 Map. Johns Hopkins Coronavirus Resource Center https://coronavirus.jhu.edu/map.html.

(2) Patick, A. K.; Potts, K. E. Protease Inhibitors as Antiviral Agents. Clin. Microbiol. Rev. 1998, 11, 614–627.

(3) De Clercq, E.; Li, G. Approved Antiviral Drugs over the Past 50 Years. Clin. Microbiol. Rev. 2016, 29, 695–747.

(4) Gorbalenya, A. E.; Baker, S. C.; Baric, R. S.; de Groot, R. J.; Drosten, C.; Gulyaeva, A. A.; Haagmans, B. L.; Lauber, C.; Leontovich, A. M.; Neuman, B. W.; Penzar, D.; Perlman, S.; Poon, L. L. M.; Samborskiy, D. V.; Sidorov, I. A.; Sola, I.; Ziebuhr, J. Coronaviridae Study Group of the International Committee on Taxonomy of Viruses. The Species Severe Acute Respiratory Syndrome-Related Coronavirus: Classifying 2019-NCoV and Naming It SARS-CoV-2. Nat. Microbiol. 2020, 5, 536–544.

(5) 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 3M from SARS-CoV-2 and Discovery of Its Inhibitors. Nature 2020, 582, 289–293.

(6) Naik, B.; Gupta, N.; Ohja, R.; Singh, S.; Prajapati, V. K.; Prusty, D. High Throughput Virtual Screening Reveals SARS-CoV-2 Multi-Target Binding Natural Compounds to Lead Instant Therapy for COVID-19 Treatment. Int. J. Biol. Macromol. 2020, 160, 1–17.

(7) Fischer, A.; Sellner, M.; Naranjan, S.; Smieško, M.; Lill, M. A. Potential Inhibitors for Novel Coronavirus Protease Identified by Virtual Screening of 606 Million Compounds. Int. J. Mol. Sci. 2020, 21, 3626.

(8) Dai, W.; Zhang, B.; Jiang, X.-M.; Su, H.; Li, J.; Zhao, Y.; Xie, X.; Jin, Z.; Peng; J.; 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, 368, 1331–1335.

(9) Anand, K.; Palm, G. J.; Mesters, J. R.; Siddell, S. G.; Ziebuhr, J.; Hilgenfeld, R. Structure of Coronavirus Main Protease Reveals Combination of a Chymotrypsin Fold with an Extra α-Helical Domain. EMBO J. 2002, 21, 3213–3224.

(10) 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.

(11) Chen, Y.; Liu, Q.; Guo, D. Emerging Coronavirus: Genome Structure, Replication, and Pathogenesis. J. Med. Virol. 2020, 92, 418–423.

(12) 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
