JCIM JOURNAL OF CHEMICAL INFORMATION AND MODELING

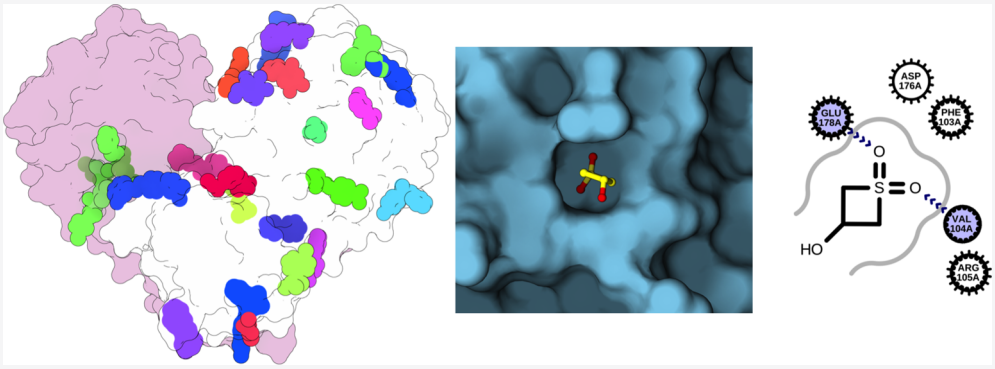
pubs.acs.org/jcim Article

Dynamic Profiling of β-Coronavirus 3CL M^{pro} Protease Ligand-Binding Sites

Eunice Cho, Margarida Rosa, Ruhi Anjum, Saman Mehmood, Mariya Soban, Moniza Mujtaba, Khair Bux, Syed T. Moin, Mohammad Tanweer, Sarath Dantu, Alessandro Pandini, Junqi Yin, Heng Ma, Arvind Ramanathan, Barira Islam, Antonia S. J. S. Mey, Debsindhu Bhowmik, and Shozeb Haider*

Cite This: *J. Chem. Inf. Model.* 2021, 61, 3058−3073 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information



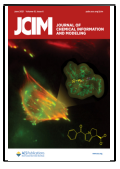
ABSTRACT: β-coronavirus (CoVs) alone has been responsible for three major global outbreaks in the 21st century. The current crisis has led to an urgent requirement to develop therapeutics. Even though a number of vaccines are available, alternative strategies targeting essential viral components are required as a backup against the emergence of lethal viral variants. One such target is the main protease (M^{pro}) that plays an indispensable role in viral replication. The availability of over 270 M^{pro} X-ray structures in complex with inhibitors provides unique insights into ligand−protein interactions. Herein, we provide a comprehensive comparison of all nonredundant ligand-binding sites available for SARS-CoV2, SARS-CoV, and MERS-CoV M^{pro}. Extensive adaptive sampling has been used to investigate structural conservation of ligand-binding sites using Markov state models (MSMs) and compare conformational dynamics employing convolutional variational auto-encoder-based deep learning. Our results indicate that not all ligand-binding sites are dynamically conserved despite high sequence and structural conservation across β-CoV homologs. This highlights the complexity in targeting all three M^{pro} enzymes with a single pan inhibitor.

INTRODUCTION

Coronaviruses (CoVs) belong to a family of positive-sense, single-stranded RNA viruses with a spherical envelope and a crownlike appearance because of their distinctive spike projections.[1,2] While α- or β-CoV infect mammals, γ- and δ-CoV can infect birds or mammals (Figure 1A).[3] Currently, seven types of CoV have been identified that infect humans, namely, human CoV 229E (HCoV-229E), OC43 (HCoV-OC43), NL63 (HCoV-NL63), Hong Kong University-1 (HCoV-HKU1), severe acute respiratory syndrome CoV (SARS-CoV), Middle East respiratory syndrome CoV (MERS-CoV), and severe acute respiratory syndrome CoV2 (SARS-CoV2).[1,2,4−9] The first four are responsible for 5−30% of common cold,[10] while the latter three cause acute lung injury, acute respiratory distress syndrome, septic shock, and multiorgan failure with a high case fatality ratio.[2,11] β-CoV alone has been responsible for three major global outbreaks in the 21st century: SARS in 2002, MERS in 2013, and COVID-19 in 2019, with a fatality rate of 10%, 34%, and 3−5%, respectively.[12]

CoV has the largest genome among any RNA viruses, with a size ranging between 26−32 kb.[13,14] The replication cycle of CoVs is initiated by the spike protein attached to the host receptor, inducing fusion events that allow viral entry into the host cell.[15] Once released inside, the viral genome is expressed into a series of proteins using multiple open reading frames (ORFs). In the SARS-CoV2 genome, 23 unannotated viral ORFs have been identified, and they include upstream ORFs that are likely to have a regulatory role: several in-frame

Received: April 19, 2021
Published: June 14, 2021



https://doi.org/10.1021/acs.jcim.1c00449
J. Chem. Inf. Model. 2021, 61, 3058−3073

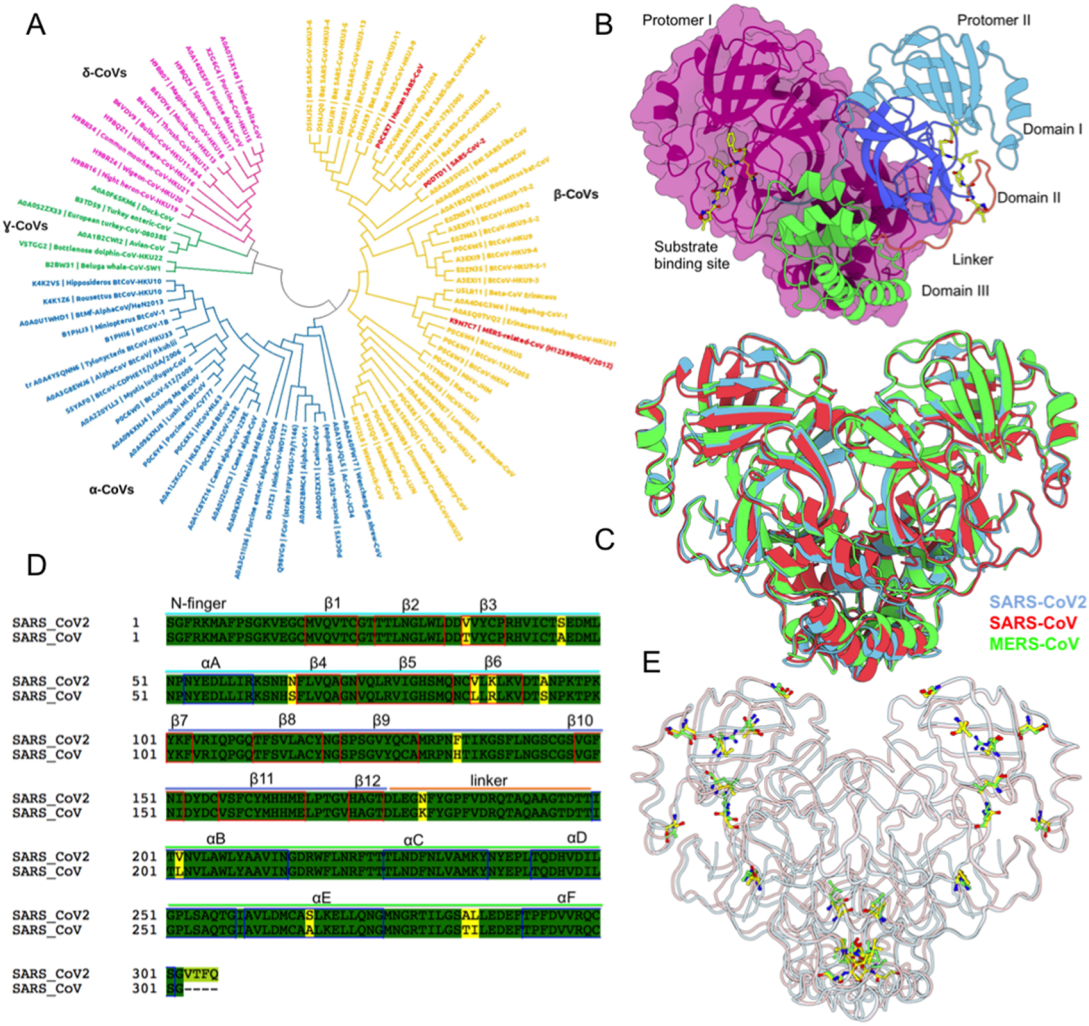


Figure 1. Overview of β-CoV 3CL M^{pro}. (A) Phylogenetic tree of the α (blue), β (yellow), γ (green), and δ (pink) CoV family; (B) structure of the dimeric SARS-CoV2 M^{pro} enzyme (PDB 6 LU7). The two protomers are represented in two different colors; the structural domains in protomer II are illustrated as cartoons; (C) comparison of SARS-CoV2 (PDB 6LU7, cyan), SARS-CoV (PDB 2C3S, red), and MERS-CoV (PDB 4YLU, green) crystal structures; (D) sequence alignment between SARS-CoV2 and SARS-CoV, highlighting the position of 12 dissimilar residues in yellow. The structural elements have been annotated on the sequence. (E) Spatial position of dissimilar residues (yellow, SARS-CoV2; green, SARS-CoV) highlighted in the M^{pro} structure.

internal ORFs within existing ORFs, resulting in N-terminally truncated products, as well as internal out-of-frame ORFs, which generate novel polypeptides.[16] Among these, two overlapping ORFs (ORF1a and ORF1b), which make up two-thirds of its genome, are translated into two large polyproteins (pp1a and pp1ab). The remaining genome is transcribed into conserved structural (spike, envelope, membrane, and nucleocapsid) and accessory proteins that are not essential for virus replication but have a role in pathogenesis.[17]

The pp1a and pp1ab polyproteins are processed by two conserved viral proteases, 3-chymotrypsin-like cysteine protease (3CL^{pro} or M^{pro}) and papain-like protease (PL^{pro}), into 16 nonstructural proteins (Nsp1−16), which are essential for viral replication and transcription.[18] M^{pro} is encoded by Nsp5 and autocleaved from polyproteins to produce a mature enzyme. The M^{pro} enzyme then cleaves 11 downstream nonstructural proteins that are important for viral replication, thereby making M^{pro} an essential protein for the viral life cycle.[19] The substrate recognition sequence of M^{pro} at most sites is x-(L/F/V)Q↓(G/A/S)-x (x = any amino acid; ↓ cleavage site), where glutamine, prior to the cleavage site, is essential.[20] No human protease with a similar cleavage specificity is known. Thus, compounds that target this cleavage site on M^{pro} will have little or no impact on human cellular proteases.[21] This makes M^{pro} an attractive drug target.

The SARS-CoV2 M^{pro} structure is a homodimer, with each protomer (residues 1−306) composed of three domains (Figure 1B). Domain I (residues 8−101) consists of 6 β-strands (β 1−6) and one α-helix (α-helix A), while domain II (residues 102−184) consists of 6 β-strands (β 7−12). The β-strands form an antiparallel β-barrel structure in each domain and uses a long linker loop (residues 185−200) to connect to domain III (residues 201−303), which has five α-helices (α-helix B-F) arranged in a compact antiparallel globular cluster.[22] The substrate-binding site is present in a cleft between domains I and II and buries the C145-H41 catalytic dyad. During the hydrolysis reaction, C145 acts as a nucleophile, while H41 acts as a base catalyst. An oxyanion hole formed by the backbone amido groups of G143 and C145 stabilizes the partial negative charge developed at the substrate cleavage bond.[23,24] The substrate-binding site consists of five subsites

https://doi.org/10.1021/acs.jcim.1c00449
J. Chem. Inf. Model. 2021, 61, 3058−3073

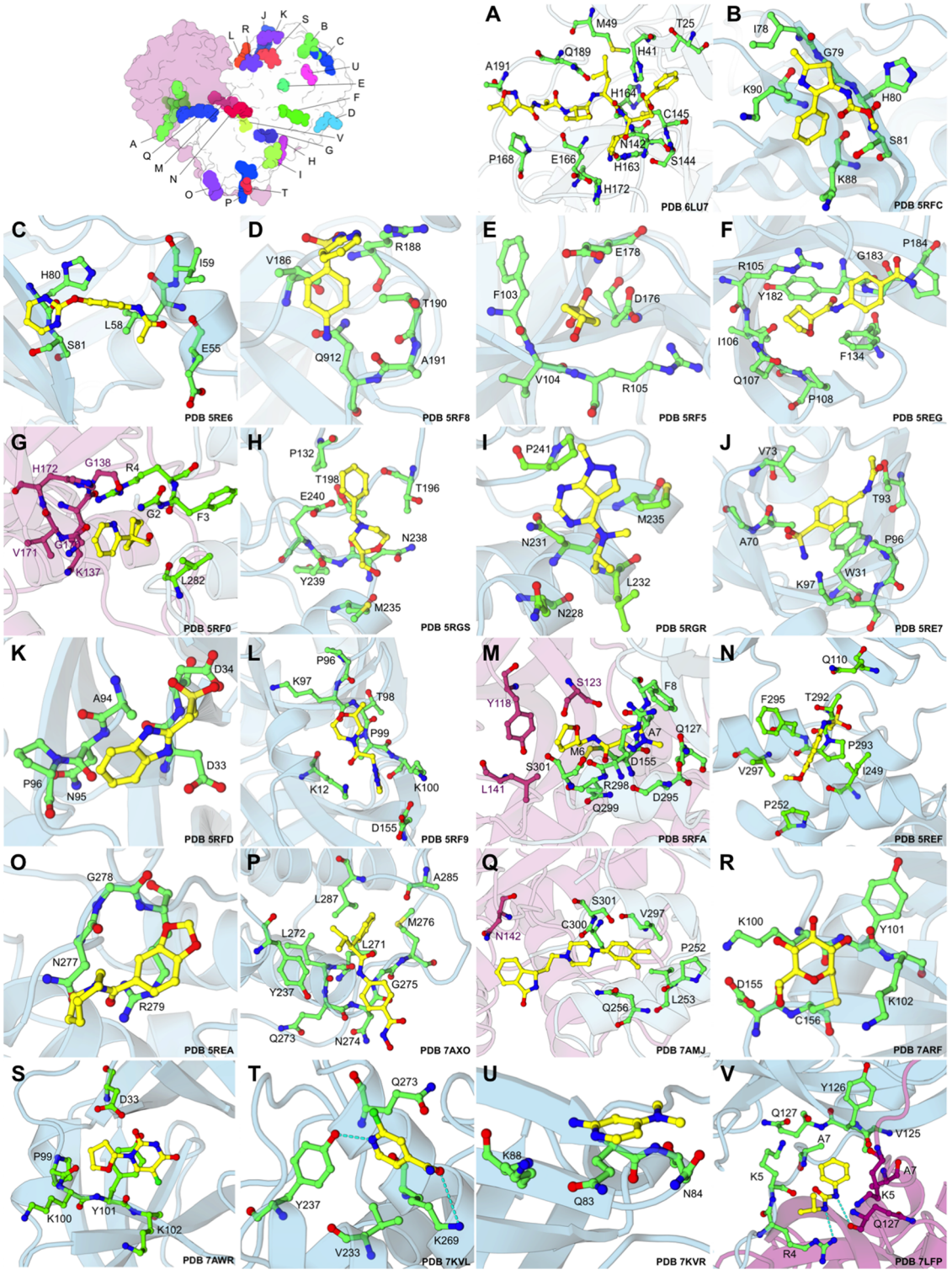


Figure 2. Ligand-binding sites in SARS-CoV2. (A) An overview of the ligand-binding sites identified from X-ray structures. While there are two copies of each binding site (one on each protomer), only one copy is illustrated. (A-V) 22 nonredundant ligand-binding sites identified from various SARS-CoV2 representative structures. Interactions between the SARS-CoV2 (green) and the ligand (yellow) in their representative ligand-binding sites. Residues within 4.0 Å of the ligand have been highlighted as green sticks. The protomers are colored in cyan and pink; the PDB entry of the representative structure is annotated in the bottom-right corner. Binding site on only one protomer is illustrated for clarity.

(S1', S1, S2, S3, and S4). Among these, subsite S1 defines enzyme specificity for glutamine in the substrate.[22] The S3 subsite is largely solvent-exposed. It lacks a typical pocket shape; however, it can be a key site where hydrophilic side chains can interact with the solvent.[25] Moreover, in the homodimer structure of the M^{pro} enzyme, the N-finger (residues 1−7) of one protomer is squeezed between domains I and II to shape the substrate-specificity pocket. This shows the importance of dimerization and N-finger orientation for substrate specificity and catalysis.[26] Further structural analysis

https://doi.org/10.1021/acs.jcim.1c00449
J. Chem. Inf. Model. 2021, 61, 3058−3073

of the SARS-CoV2 M^{pro} enzyme showed that domains I and II are connected via seven residues (D92-P99) that contribute to the substrate-binding site.[21,22] Domain III contributes to the proteolytic activity via dimerization of the M^{pro} enzyme.[22] Dimerization is important because monomeric M^{pro} does not exhibit any catalytic activity.[27] Because M^{pro} is a symmetric homodimer, two copies of ligand-binding sites are present, one on each protomer.

A comparison of the SARS-CoV2, SARS-CoV, and MERS-CoV M^{pro} sequences revealed that SARS-CoV2 is 96% similar to SARS-CoV and 51% similar to MERS-CoV. A structural superimposition of all the M^{pro} enzymes displayed an overall root-mean-squared deviation (RMSD) of 0.85 Å (± 0.16 Å), with a very high degree of structural conservation around the catalytic dyad in the substrate-binding site, suggesting very similar substrate recognition profiles among these proteins (Figure 1C). The difference between SARS-CoV2 and SARS-CoV M^{pro} is 12 amino acids in each protomer (Figure 1D,E).

The past year has seen a dramatic progress in SARS-CoV2 research (covid19primer.com). Significant efforts have gone into the design of M^{pro} inhibitors that target the substrate-binding pocket.[21,22,28−30] This also includes the inhibitor design via various in silico methods.[31−36] Recent progress in M^{pro} inhibitors has been reviewed elsewhere.[37−39] Of considerable note is the COVID Moonshot project, which generates data via open science discovery of M^{pro} inhibitors by combining crowdsourcing, high-throughput experiments, computational simulations, and machine learning.[40] This alone has generated over 258 structures of fragments and leadlike molecules in complex with the M^{pro} protease (www.covid.postera.ai/covid). Other large-scale efforts using the crystallographic screening of fragments and drug repurposing libraries have identified allosteric drug-binding sites.[41,42]

Over 270 crystal structures of SARS-CoV2 M^{pro} are present in the protein data bank (PDB), including apo- and cocomplexes with inhibitors (Table S1). Additional similar data are also available on SARS-CoV (Table S2) and MERS-CoV (Table S3) M^{pro} enzymes. However, to date, no comprehensive, consolidated, comparison of ligand-binding sites and their complexes determined by X-ray crystallography has been reported. In this study, we map nonredundant ligand-binding sites from all crystal structures of SARS-CoV2 M^{pro} available in the PDB. Then, we perform 25 μs of adaptive molecular dynamics (MD) simulations on the apo M^{pro} structures of SARS-CoV2, SARS-COV, and MERS-CoV and investigate the structural conservation of the ligand-binding sites using Markov state models (MSMs). We annotate each binding site with a measure of correlated evolution at the residue level. Furthermore, we explore the differences in the conformational dynamics of the M^{pro} enzymes using a deep-learning approach, namely, a variational autoencoder (AE) with convolutional filters (CVAE).[43] Our results highlight that even though with a structural overlap of <1 Å, the conformational dynamics of SARS-CoV2, MERS-CoV, and SARS-CoV are very different. A persistence analysis and comparison of the structural conservation of the ligand-binding sites in β-CoV homologs highlight the complexity in targeting all three M^{pro} enzymes with a single pan inhibitor.

RESULTS

Mapping the Binding Sites. The PDB was searched for β-CoV M^{pro} entries. A total of 271 SARS-CoV2 structures were identified. Out of these, there were 38 structures with no ligands, and they were excluded from any further study. The remaining 233 structures were downloaded for a detailed structural analysis and are listed in Tables S1−S6. The key interacting residues between the inhibitors and M^{pro} were mapped (Figure 2; Figure S1). In total, 22 different binding sites were identified. These have been labeled A−V in Figure 2 and are listed in Table 1. A detailed structural description of the binding site is provided in the Supporting Information. Site A is the substrate-binding or the active site. It is worth emphasizing that not all binding sites (B−V) are allosteric in nature. Only some ligands that are bound to site Q and N showed allosteric inhibition.[42] Many of these ligands are small fragments that are bound to neither an allosteric nor an active site, and even up to 100 μM, they did not show any antiviral activity.[42]

Table 1. SARS-CoV2 M^{pro} Ligand-Binding Sites[a]

binding site	binding-site residues	ligand ID	PDB (representative structure)	number of ligands
A	T25, H41, M49, Q189, H164, A191, C145, N142, S144, H163, E166, H172, P168	N3	6 LU7	185
B	I78, G79, H80, S81, K88, K90	K1Y	5RFC	5
C	H80, I59, E55, L58, S81	O0S	5RE6	2
D	V186, R188, T190, A191, Q912	SFY	5RF8	1
E	F103, E178, D176, R105, V104	HV2	5RF5	4
F	R105, Y182, G183, P184, F134, P108, Q107, I106	LWA	5REG	1
G	H172, G138, R4, G2, F3 L282, K137, G170, V171	T5D	5RF0	1
H	P132, T198, Y196, E240, Y239, M235, N238	S7V	5RGS	4
I	P241, M235, I232, N231, N228	K1G	5RGR	1
J	A70, V73, T93, P96, W31, K97	T0S	5RE7	1
K	P96, N95, A94, D34, D33	T6J	5RFD	3
L	P96, T98, P99, K100, D155, K12, K97	S7D	5RF9	3
M	Y118, L141, S123, F8, Q127, D295, A7, D155, R298, Q299, M6	JGY	5RFA	2
N	Q110, F295, V297, T292, P293, I249, P252	6SU	5REF	2
O	N27, G278, R279	JGP	5REA	1
P	L287, A285, M276, G275, N274, L271, Q273, Y237, L272	QCP	7AXO	2
Q	N142, C300, S301, V297, P252, L253, Q256	RMZ	7AMJ	5
R	K100, Y101, K102, C156, D155	UHG	7ARF	6
S	D33, K102, Y101, K100, P99	RVW	7AWR	2
T	V233, Y237, K269, Q273	X4P	7KVL	1
U	Q83, N84, K88	X4V	7KVR	1
V	R4, K5, A7, V125, Y126, Q127	XY4	7LFP	1

[a]The number of ligands represents unique chemical entities that bind to the particular ligand-binding site.

Structural Dynamics of the β-CoV M^{pro} Enzymes. To further understand the structural dynamics of the β-CoV M^{pro}

https://doi.org/10.1021/acs.jcim.1c00449
J. Chem. Inf. Model. 2021, 61, 3058−3073

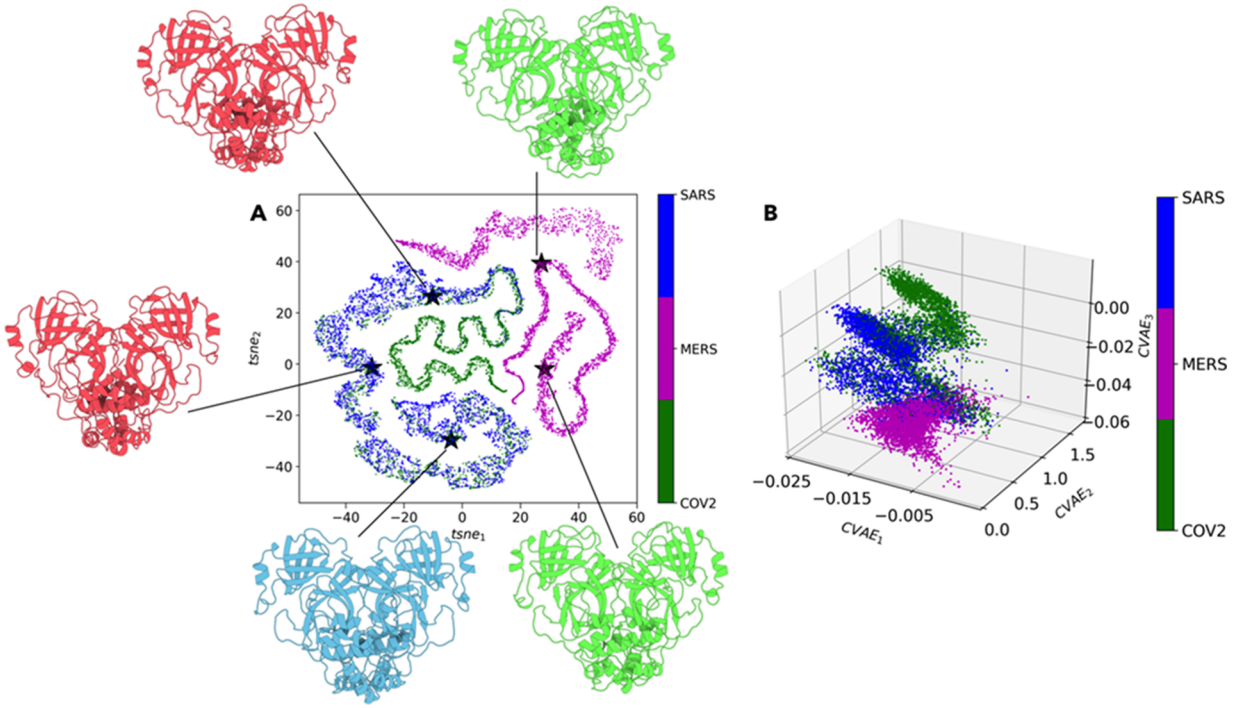


Figure 3. CVAE-based deep-learning analysis. Low-dimensional latent space of CVAE-learned features of the high-dimensional input in (A) 2D representation and in (B) 3D representation. Original high-dimensional data are transformed into a distance matrix format, which is then fed to the CVAE architecture. The CVAE captures the intrinsic features of the high-dimensional data that are necessary to describe the original system behavior. This captured information is then shown in the 3D format (right) and in the 2D format (left), following the t-sne treatment. The centroids, as detected by k-means algorithms, are illustrated. The results show that MERS-CoV (green) dynamics is very different from SARS-CoV (magenta) or SARS-CoV2 (blue) dynamics.

enzymes, MSM-based adaptive sampling MD simulations were conducted. These simulations have an advantage over classical MD in exploring under sampled states without a predetermined bias. The sampling and analysis mainly focused on investigating the differences in the dynamics of the M^{pro} enzyme ligand-binding sites. SARS-CoV2 and SARS-CoV are 96% identical, with a difference of only 24 residues out of 612. When structurally aligned, the RMSD of the protein backbone is 0.61 Å. This is also similar to when comparing with the MERS-CoV structure, where the sequence similarity is ∼51% and the RMSD of the structural alignment is 0.51 Å.

The conformational drift during the course of the simulations was assessed using Cα RMSD. Conventional RMSD fitting methods fail to separate the regions of different stabilities. To resolve this issue, we used a fraction (%) of the Cα atoms for the alignment. Beyond this fraction, there is a sharp increase in the RMSD value for the remainder of the Cα atoms. At 60%, the core of the M^{pro} could be superimposed to less than 1.2, 1, and 0.9 Å for SARS-CoV, SARS-CoV2, and MERS-CoV M^{pro} structures (Figure S2A), respectively. The Cα atoms above 60% cutoff predominantly belong to dimerization domain III, the linker loop, and the loops in domains I and II. The antiparallel β-barrel structures show the least deviation (Figure S2B,D).

CVAE-Based Deep-Learning Analysis. To further resolve the differences in the collective conformational fluctuations between the M^{pro} simulations, a CVAE was used (Figure S3). The CVAE can completely cluster the three different β-CoV M^{pro} types based on the local and global conformational dynamics (Figure 3). Here, SARS-CoV2 and SARS-CoV behave similar to each other, while the MERS-CoV behavior is very different. It should be noted that clustering using traditional features such as root mean square fluctuation (RMSF), root mean square deviation (RMSD) or native contacts were unable to distinguish differences in dynamics among these three types of closely related β-CoV homologs (Figure S4), proving the sensitivity of the CVAE implementation. The details of the CVAE analysis are described in the Supporting Information.

Markov State Model. The main aim for building an MSM was to investigate how the various binding sites identified from X-ray crystallography (Tables 123) were linked dynamically in the network of metastable states and transition probabilities among them. The choice of this method was based on the ability of MSM methods to use large ensembles of short-timescale trajectories for sampling events that occur in slow timescales.[44,45] The metastable states are an ensemble of structural conformations that interconvert quickly within the ensemble and slowly between them. These ensembles broadly correspond to the different basins on the free energy landscape

Table 2. SARS-CoV M^{pro} Binding Sites

binding site	binding-site residues	ligand ID	PDB	number of ligands[a]
A	T25, L27, H41, V42, T45, A46, M49, F140, L141, N142, G143, S144, C145, H163, H164, M165, E166, H172, V186, Q189, Q192	D03	5 N19	44
F	H134, P184, G183, Y182, F181, R105	MES	2V6N	1

[a]The number of ligands represents unique chemical entities that bind to the particular ligand-binding site.

https://doi.org/10.1021/acs.jcim.1c00449
J. Chem. Inf. Model. 2021, 61, 3058−3073

Table 3. MERS-CoV M^{pro} Binding Sites

binding site	binding-site residues	ligand ID	PDB	number of ligands[a]
A	H41, F143, L144, C145, G146, S147, C148, H166, Q167, M168, E169, A171, H175, Q192	QZG	6VGZ	10

[a]The number of ligands represents unique chemical entities that bind to the particular ligand-binding site.

(FEL). MSMs provide a powerful method for detecting metastable states as well as calculating kinetics and free energies by integrating any number of simulations into a single statistical model.[44−49]

We first used φ and ψ dihedral angles of the 24 residues that are dissimilar between SARS-CoV2 and SARS-CoV (Figure 1D) as the input data. However, these data were not sufficient to build a converged MSM. We then included φ and ψ dihedral angles of all residues and the $\chi 1$ angle from the 24 residues that were different as input data to construct the MSM. The dimensionality of the data was further reduced through time-lagged independent component analysis (tICA) and the models built using the PyEMMA software, from a set of 500 short 50 ns enhanced sampling MD simulations. It was possible to build a converged MSM with a lag time of ≥10. Shorter lag times provide more structural details but can underestimate the populations of important states, while simulations with longer lag times provide better population estimates but obscure intermediate states. The data were clustered into 100 microstates, and their distribution on the FEL is presented in Figures S5−7. Transition pathways were then generated to identify metastable conformations. In total, 5 metastable states were identified for SARS-CoV2, 5 for SARS-CoV, and 4 for MERS-CoV (Figure 4).

Dynamic Pocket Tracking. Because many of the ligand-binding sites appear together on the M^{pro} surface, we investigated the spatiotemporal evolution of the binding pockets. The protein conformations of the metastable states were searched for the presence of the experimentally reported binding sites. The site was described as open, if it could hold a minimum of five water molecules, which was a coarse equivalent of a small fragment. A comparison of equivalence was then made between the sites identified from the simulation data and those from crystallographic experiments. A comprehensive list of the binding sites and their persistence across metastable states identified from SARS-CoV2, SARS-CoV, and MERS-CoV M^{pro} dynamics is presented in Table 4.

Sites A−L, P, R, S, and V are present in all metastable states in SARS-CoV2. Based on the evolutionary conservation scores, most of the pockets (except F, J, K, N, O, and U) are more conserved than the surface residues, with the strongest evolutionary signal observed for pockets B, P, R, and T (Figure S8).

Two copies (one in each protomer) of site M are present in state 5, only one in states 2 and 3, and none in states 1 and 4. In the crystal structure (PDB 5RFA), the carboxylic acid side chain of D295 makes interactions with the hydroxyl group side chain of T111; the side chain of Q299 forms a hydrogen bond with the backbone carbonyl oxygen atom of R4 in the N-finger; and the guanidinium side chain of R298 forms a hydrogen bond with the backbone carbonyl oxygen atom of I152. These interactions lock α-helix F in domain III to antiparallel β-barrel in domain II. The ligand occupying the large cavity at the interface further helps stabilize the local structural elements around site M. In the absence of the ligand in the binding site and because of dynamic fluctuations, the R298-I152 interaction is lost. The side chain of R298 is free to rotate and can adopt a conformation that can occupy the empty binding site (Figure 5A). Furthermore, the C-terminal tail also occludes one of the binding sites in states 2 and 3.

Site N is a deep cleft between α-helix D and F and is spatially positioned adjacent to site M. The side chain of F294 (α-helix F) is shared between both the sites. The rotation of the phenyl side chain controls the opening and closure of site N. When site N is open, the phenyl side chain of F294 is positioned on α-helix F. In the closed state, the F294 side chain is positioned in the cleft. This conformation is analogous to that observed when a ligand is bound to site M. Site N is also conjoined with another larger cavity that runs orthogonal to it. When the ligand binds to this pocket (as in PDB 7AGA), the conformation of the side chain of F294 is similar to that observed in site M, which occludes site N. We observe all these conformations of F294 in our metastable states. The N site is present in both protomers in state 1 and in one of the two protomers in states 2, 3, 4, and 5. The orthogonal site is present in conjunction with site N in at least one of the protomers (Figure 5B).

Site O is a pseudoligand-binding site on the loop between α-helix E and F. When bound, the ligand is completely solvent-exposed and interacts with the protein structure by creating hydrogen bonds with the side chains of N277 and R279. These interactions stabilize the flexibility of this loop. In the simulated apo structure, when the ligand is absent, this loop is highly mobile, and the side chains of N277 and R279 display enhanced flexibility (Figure 5C). This results in the loss of the conformation of the loop to which the ligand binds. The conformation of the loop, similar to that adopted in the representative structure, is not observed in any metastable state.

Site Q at the interface between the two protomers is spatially positioned between the distal ends of α-helices B, D, and F. At the end of α-helix F is a short C-terminal tail (residues 300−306). In the representative crystal structure (PDB 7AMJ), the tail orients away from the α-helical dimerization domain III and is sandwiched at the interface between domain II of both protomers, away from where the ligand binds. This provides enough space for the ligand to position in the binding site Q. During the SARS-CoV2 M^{pro} apo simulations, the C-terminal tail displays dynamic flexibility and can adopt multiple conformations. In addition to the conformation observed in the representative structure, one of the adopted conformation of the loop occludes the binding site Q and prevents any ligand binding (Figure 5D). This conformation is similar to that observed in the PDB 6 LU7 structure. Site Q is present between one interface in states 3 and 4, completely occluded in states 1 and 2, and is present at both interfaces in state 5.

In the representative structure of site T (PDB 7KVL), the fragment forms hydrogen bonds with the hydroxyl group of Y237 (α-helix C) and the side chain of K269 (α-helix E). In the apo simulation, when the ligand is absent, the side chain of these residues can occupy the space where the fragment binds (Figure 5E). This results in the loss of this site in states 1, 2, and 5. However, the site is present in both protomers in state 4 and in only one protomer in state 3.

Site U is a solvent-exposed pseudoligand-binding site that is stabilized by the hydrogen bond interaction between the side

https://doi.org/10.1021/acs.jcim.1c00449
J. Chem. Inf. Model. 2021, 61, 3058−3073

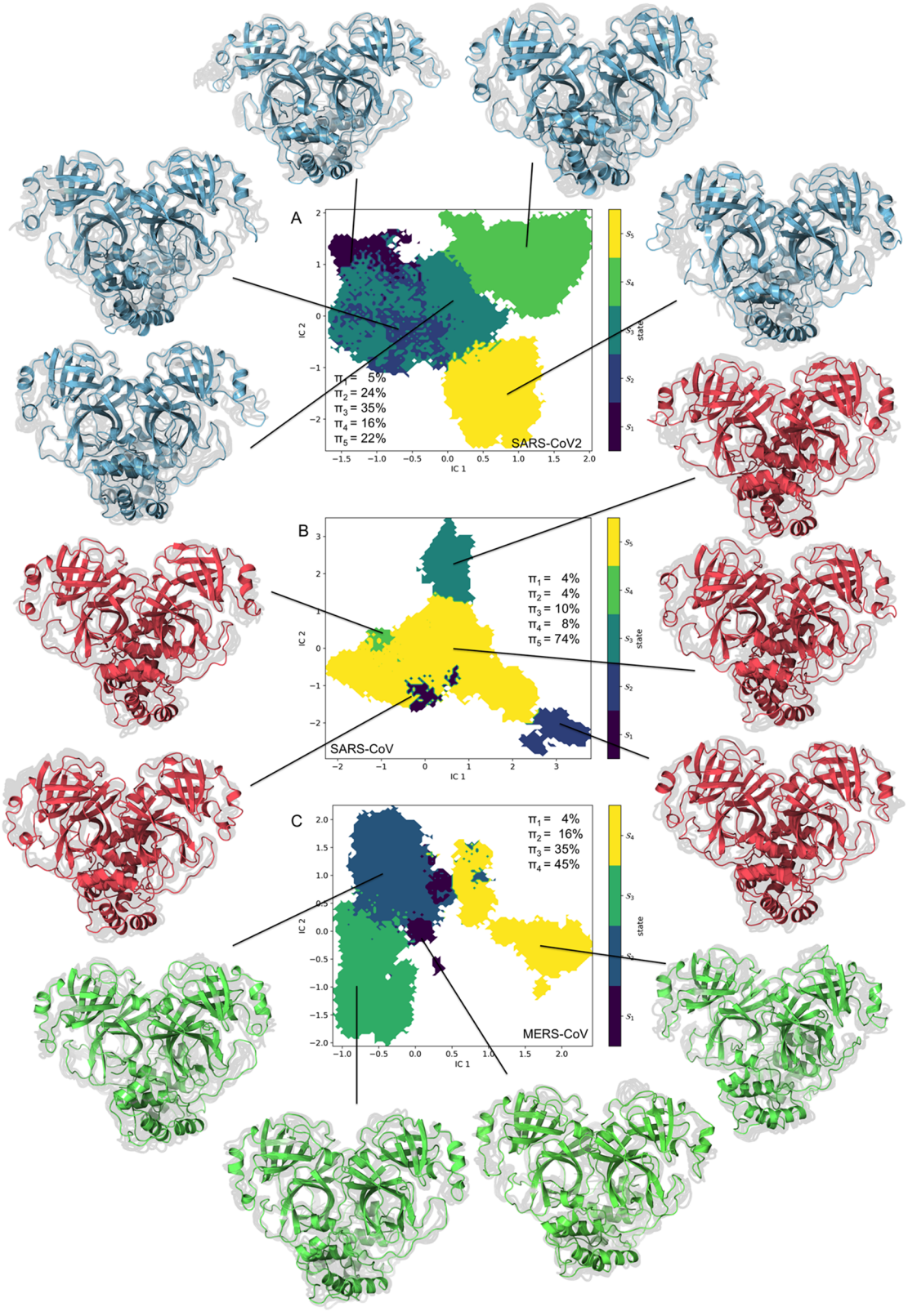


Figure 4. Markov State Network. Macrostate distributions of (A) SARS-CoV2, (B) SARS-CoV, and(C) MERS-CoV conformations projected onto the first two time-lagged independent components (ICs). The population of each state (π) is indicated in the figure. The state with the highest population is classified as the dominant state. The representative metastable structures are illustrated in Figures S5−S7.

chains of K88 and Q83. A part of this binding site is formed by N84, which is present in the loop between β5-β6 strands. In the absence of the fragment, the residues from this site can adopt multiple conformations, which would be unsuitable for the stacking of any fragment in this site (Figure 5F). The conformation of residues comparable to the representative site is present in both protomers in states 1 and 3, as well as in one protomer in states 2, 4, and 5.

https://doi.org/10.1021/acs.jcim.1c00449
J. Chem. Inf. Model. 2021, 61, 3058−3073

Table 4. Dynamic Tracking of Ligand-Binding Sites[a]

Site	Representative PDB Binding Site	SARS-CoV2 Metastable states				
		1	2	3	4	5
A	6LU7	x	x	x	x	x
B	5RFC	x	x	x	x	x
C	5RE6	x	x	x	x	x
D	5RF8	x	x	x	x	x
E	5RF5	x	x	x	x	x
F	5REG	x	x	x	x	x
G	5RF0	x	x	x	x	x
H	5RGS	x	x	x	x	x
I	5RGR	x	x	x	x	x
J	5RE7	x	x	x	x	x
K	5RFD	x	x	x	x	x
L	5RF9	x	x	x	x	x
M	5RFA	-	-x	-x	-	x
N	5REF	x	-x	-x	-x	-x
O	5REA	-	-	-	-	-
P	7AXO	x	x	x	x	x
Q	7AMJ	-	-	-x	-x	x
R	7ARF	x	x	x	x	x
S	7AWR	x	x	x	x	x
T	7KVL	-	-	-x	x	-
U	7KVR	x	-x	x	-x	-x
V	7LFP	x	x	x	x	x

Site	Representative PDB Binding Site	SARS-CoV Metastable states				
		1	2	3	4	5
A	6LU7	x	x	x	x	x
B	5RFC (T35 R88)	x	x	x	x	x
C	5RE6	x	x	x	x	x
D	5RF8	x	x	x	x	x
E	5RF5 (K180)	x	x	x	x	x
F	5REG (H134)	x	x	x	x	x
G	5RF0	-	-	-	-	-
H	5RGS	x	x	x	x	x
I	5RGR	x	x	x	x	x
J	5RE7 (S94)	x	x	x	x	x
K	5RFD (S94)	-	-	-	-	-
L	5RF9	x	x	x	x	x
M	5RFA	-	-	-	-	-
N	5REF	x	-	x	-	x
O	5REA	-	-	-	-	-
P	7AXO	x	x	x	x	x
Q	7AMJ	x	x	x	x	x
R	7ARF	x	x	x	x	x
S	7AWR	x	x	x	x	x
T	7KVL	-x	-	-x	-x	-
U	7KVR (R88)	-x	-x	x	x	-
V	7LFP	x	x	x	x	x

Site	Representative PDB Binding Site	MERS-CoV Metastable states			
		1	2	3	4
A	6LU7	x	x	x	x
B	5RFC	x	x	x	x
C	5RE6	x	x	x	x
D	5RF8	-x	-x	-x	x
E	5RF5	-x	x	x	-x
F	5REG	x	x	x	x
G	5RF0	-x	-x	-x	-
H	5RGS	-	-	-	-
I	5RGR	x	-x	-x	-x
J	5RE7	-	-	-	-
K	5RFD	x	x	x	x
L	5RF9	x	x	x	x
M	5RFA	x	x	x	x
N	5REF	-	-	-	-
O	5REA	-	-	-	-
P	7AXO	x	x	x	x
Q	7AMJ	x	x	x	x
R	7ARF	-	-	-	-
S	7AWR	x	x	x	x
T	7KVL	-x	x	-x	-x
U	7KVR	-	-	-	-
V	7LFP	x	x	x	x

[a]The persistence of the ligand-binding sites in (left) SARS-CoV2, (middle) SARS-CoV, and (right) MERS-CoV metastable states after comparison with the representative X-ray structures. Residues in SARS-CoV that are different from SARS-CoV2 are highlighted in parenthesis. Binding sites that are present in both protomers in the metastable state and in the representative X-ray structure are indicated by an "x" sign, those that are absent are noted by a "-", and those that are present in at least one protomer are denoted by a "-x" sign.

In SARS-CoV, sites equivalent to A-F, H-J, P, R, S, and V are present in all metastable states. These sites are well-defined pockets and are comparable to the X-ray crystal structures of SARS-CoV2 (Table 4).

Site G, which is formed at the interface of the two protomers, is lost in all metastable states of SARS-CoV. During the dynamics of the apo state, the loop between β8-β9 becomes flexible. The mobility of the loop pushes the N-finger, which is tucked below the substrate-binding A site to collapse on site G (Figure 5G). In this conformation, no ligand would be able to bind to this site.

Binding site K is also lost in all metastable states in SARS-CoV. In this site, the hydroxyl group side chain of S94 is present ($V94_{SARS\text{-}CoV2}$). In the absence of the ligand, the side chains of D33 and S94 orient toward each other, where they form a hydrogen bond. This stable interaction is spatially positioned on the site where the ligand binds (Figure 5H), thus completely obstructing the binding site.

Unlike in SARS-CoV2, the equivalent site on SARS-CoV, where the ligand binds in site M is absent in all metastable states. In the representative site, the side chain of R298 forms a hydrogen bond with the backbone oxygen of I152. During the simulation, this interaction is lost, and the side chain of R298 in α-helix F becomes flexible and can adopt multiple conformations. One such conformation blocks the ligand-binding pocket M. The dynamics observed in this pocket are similar to that observed in SARS-CoV2 simulations (Figure 5A).

The dynamic behaviors of residues in sites N, O, T, and U are also similar to that observed in SARS-CoV2 (Figure 5B,C,E,F). The formation or the dissolution of site N depends upon the conformation of the phenyl side chain in F294. The site is present when the side chain orients away from the binding site and is absent when the side chain is positioned toward the binding site. Site N is observed in states 1, 3, and 5, while it is absent in states 2 and 4. Site O is a pseudo-binding site present on a highly dynamic loop. In the apo state, the N277-G278-R279 loop is highly flexible. This permits the side chains to adopt multiple conformations. However, none of the conformations are structurally similar to that which binds the ligand in the representative structure. The SARS-CoV structure lacks a C-terminal tail (PDB 2C3S); hence, site Q is always present in the dynamic structures. The presence of site T depends on the conformation of Y237, Q273, and K269 side chains. In the absence of the fragment, the side chains are dynamics and can occlude the binding site. Site T is present in one protomer in states 1, 3, and 4 and is absent in states 2 and 5. The dynamics of residues in site U, where R88 replaces K88, are similar to that observed in SARS-CoV2. The side-chain conformation of residues on which the fragment stacks is observed in states 1, 2, 3, and 4 and is absent in state 5.

From the list of 12 residues that are dissimilar between SARS-CoV2 and SARS-CoV (Figure 1D) in each protomer, V35 and K88 (backbone) are present in site B. Equivalent residues in SARS-CoV are T35 and R88, respectively. These residues have similar sizes and therefore do not alter the dimensions of the binding site. However, a change from $V35_{SARS\text{-}CoV2}$ to $T35_{SARS\text{-}CoV}$ does alter the surface charge pattern around the binding site. The side chain of $K88_{SARS\text{-}CoV2}$ ($R88_{SARS\text{-}CoV}$) contributes toward stabilizing fragment binding in site U, where it makes a hydrogen bond with Q83. $N180_{SARS\text{-}CoV2}$ is replaced with a $K180_{SARS\text{-}CoV}$ at the entrance

https://doi.org/10.1021/acs.jcim.1c00449
J. Chem. Inf. Model. 2021, 61, 3058−3073

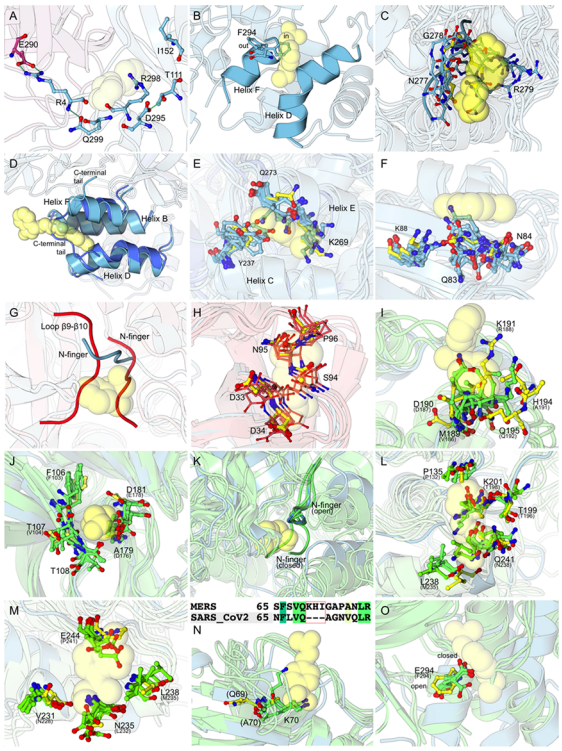


Figure 5. Lost ligand-binding sites on (A-F) SARS-CoV2, (G-H) SARS-CoV, and (J-O) MERS-CoV M^{pro}. (A) Site M; when the interaction between R298-I152 is lost, the R298 side chain becomes flexible and obstructs the ligand-binding site. (B) Site N; the side chain of F294 exists in two conformations. When facing inward, it occludes the binding site. (C) Site O is a part of a dynamic loop, which is unable to maintain the structure to which the ligand binds. Conformations of the loop from all metastable states are illustrated. (D) Site Q; in the absence of the ligand, the C-terminal tail collapses in the binding site and blocks it. Conformation of the helices from the representative PDB (id 7AMJ, cyan) and that of state 3 (dark blue) have been highlighted. (E) Site T; the side chains of Y237, K269, and Q273 (from all states) in the absence of the fragment can occupy the binding site. (F) Site U; the flexible side chains of Q83 and N84 (from all states) can disrupt the conformation on which the fragment stacks. (G) Site G; formed at the interface when the N-finger is tucked below the substrate-binding site A. Structural changes in loop β9-β10 destabilize the N-finger, which results in its collapse on the ligand-binding site. The position of the N-finger in the representative structure (PDB 5RF0) is colored cyan, and the conformation of state 4 is shaded in red. (H) Site K; the side chains of S94 and D33 can form a hydrogen bond, which occludes the space where the ligand binds in the representative structure (PDB 5RFD). (I) Site D; the longer side chain of M189 (in place of $V186_{SARS\text{-}CoV2}$) obstructs the binding site. (J) Site E; loss of steric repulsion prevents this site to stay perpetually open; (K) Site G; the N-finger collapses on the ligand-binding site as a result of the fluctuations in the β9-β10 loop. The position of the N-finger in the representative structure (PDB 5RF0) is colored in cyan, and the conformations of all metastable states are shaded in green. (L) Site H; the longer lysyl side chain of K201 ($T198_{SARS\text{-}CoV2}$) blocks the binding site. (M) Site I; the side chain of E244 ($P241_{SARS\text{-}CoV2}$) occludes the binding site; (N) Site J; an insertion of three residues at position 70 increases the length of the loop between β4-β5. The presence of a larger K70 side chain and the conformation of the loop restrict the dimensions of the binding site; (O) Site N; the side chain of E294 is always in the closed conformation and impedes the binding site. The representative structure is represented in cyan and yellow spheres that indicate the spatial position of where the ligand binds in the corresponding representative structure.

of binding site E. This alters the surface charge around the entrance of the binding site E toward a more positive charge. Both residues, in their respective proteins, orient toward the solvent and do not interact with any other part of the protein. The backbone atoms of $A94_{SARS\text{-}CoV2}$ ($S94_{SARS\text{-}CoV}$) form the boundary of the binding site J, while the side chain contributes to binding site K. The side-chain interaction between S94 with D33 occludes the pocket and in turn has an effect on the conformation of the binding site. $F134_{SARS\text{-}CoV2}$ is replaced with $H134_{SARS\text{-}CoV}$ in site F. A protonated histidine side chain at the ϵ-nitrogen atom can form strong interactions with the ligand in SARS-CoV. $V202_{SARS\text{-}CoV2}$ is positioned at the start of helix B and is a part of the large channel-like cavity between domains II and III. Ligand AT7519 (PDB 7AGA) binds in this cavity. A deep cleft branches off this channel and forms site N. A change from $V202_{SARS\text{-}CoV2}$ to $L202_{SARS\text{-}CoV}$ slightly decreases the dimensions of this channel. The backbone $A285_{SARS\text{-}CoV2}$ and the side chain of $L286_{SARS\text{-}CoV2}$ form the boundary of the P site. A change to $T285_{SARS\text{-}CoV}$ and $I286_{SARS\text{-}CoV}$ does not alter the dimensions of the binding site; however, these residues have been implicated in being involved in cooperative effects and enhancing dimerization in SARS-CoV.[50] The hydroxyl side chain of $S46_{SARS\text{-}CoV2}$ ($A46_{SARS\text{-}CoV}$) orients near the edge of the substrate-binding subsite S2. Similarly, residue 65 ($N65_{SARS\text{-}CoV2}$ and $S65_{SARS\text{-}CoV}$) is positioned near a cavity at the entrance of the antiparallel β-barrel in domain I, which is a potential binding site. However, we could not find any ligand that interacts with S46 or N65. Residues $V_{SARS\text{-}CoV2}$/$L_{SARS\text{-}CoV}$86 and $S_{SARS\text{-}CoV2}$/$A_{SARS\text{-}CoV}$267 are located in the core of the enzyme and do not contribute to any cavities identified on SARS-CoV2 or SARS-CoV.

In MERS-CoV, sites A-C, F, K-M, P, Q, S, and V are present in all metastable states. Site D is present in both protomers in state 4 and in one protomer in states 1, 2, and 3. Of particular note is the substitution of $M189_{MERS\text{-}CoV2}$ (in place of $V186_{SARS\text{-}CoV2}$) in this site. The longer side chain of M189 obstructs the ligand-binding site in some states (Figure 5I).

Site E is present in both protomers in states 2 and 3 and in one protomer in states 1 and 4. In SARS-CoV2, the side chains of D176 and E178 form the boundary of this site. In the apo state, the charge repulsion between the two negatively charged side chains prevents the closure of this site in SARS-CoV2. However, $D176_{SARS\text{-}CoV2}$ is replaced with $A179_{MERS\text{-}CoV}$ and $E178_{SARS\text{-}CoV2}$ with $D181_{MERS\text{-}CoV}$. In the absence of the ligand, and with no charge repulsion between the negatively charged side chains, the side chain of D181 obstructs the binding site in some metastable states (Figure 5J).

Site G is present in one protomer in states 1, 2, and 3 and is absent in state 4. In SARS-CoV2, the N-finger is tucked below site A, which provides enough space at the interface for the ligand to bind in site G. In the simulated apo state of MERS-CoV and similar to that observed in SARS-CoV2, the N-finger can also collapse and occupy the binding site, resulting in its closure (Figure 5K).

In Site I, $N228_{SARS\text{-}CoV2}$, $L232_{SARS\text{-}CoV2}$, $M235_{SARS\text{-}CoV2}$, and $P241_{SARS\text{-}CoV2}$ are replaced with $V231_{MERS\text{-}CoV2}$, $N235_{MERS\text{-}CoV2}$, $L238_{MERS\text{-}CoV2}$, and $E244_{MERS\text{-}CoV2}$, respectively. The longer carboxylic side chain in $E244_{MERS\text{-}CoV}$ can adopt a conformation that obstructs the binding site (Figure 5L). This is observed in at least one protomer in states 2, 3, and 4, while the binding site is clear in state 1.

Site T is present in both protomers in state 2 and in one protomer in states 1, 3, and 4. Here, $Y273_{MERS\text{-}CoV}$ in

https://doi.org/10.1021/acs.jcim.1c00449
J. Chem. Inf. Model. 2021, 61, 3058−3073

substituted in place of $L275_{SARS\text{-}CoV2}$. Furthermore, a large indole ring in $W236_{MERS\text{-}CoV}$ replaces the smaller side chain of $V233_{SARS\text{-}CoV2}$, making the binding site shallower than its representative structure. Taken together, the side chains of $W236_{MERS\text{-}CoV}$ and $Y273_{MERS\text{-}CoV}$ act like a wedge to split and widen α-helices C and E. Therefore, site T is persistently more open when compared with the dynamics of SARS-CoV2 or SARS-CoV.

Sites H, J, N, O, R, and U are absent in all metastable states in MERS-CoV. In site H, the substitution of a shorter hydroxyl group in $T198_{SARS\text{-}CoV2}$ to a longer lysyl side chain in $K201_{MERS\text{-}CoV}$ completely obstructs the binding sites in all metastable states (Figure 5M). In MERS-CoV, loop $\beta4$-$\beta5$ is extended by the insertion of three residues between positions 69−70. As a result, there is a change from $A70_{SARS\text{-}CoV2}$ to a lysine at this position. The longer lysyl side chain obstructs site J where the ligand binds (Figure 5N). $F294_{SARS\text{-}CoV2}$ is substituted with $E294_{MERS\text{-}CoV}$ in site N. Unlike in SARS-CoV2 and SARS-CoV, the side chain of $E294_{MERS\text{-}CoV}$ points toward the N site cleft, which blocks site N (Figure 5O). Ligands interact with site S by forming a disulfide bond with $C156_{SARS\text{-}CoV2}$. However, in MERS-CoV, the cysteine residue is replaced with $V159_{MERS\text{-}CoV}$, which would prevent any disulfide bond formation. In site U, the side chain on which the ligand stacks is absent because of the substitution of $N84_{SARS\text{-}CoV2}$ by $G87_{MERS\text{-}CoV2}$.

DISCUSSION

Despite tremendous advances in the inhibitor design for SARS-CoV2 M^{pro} enzymes, our understanding of the role of the structural dynamics of the experimentally identified ligand-binding sites remains largely uncharacterized. Most MD studies have focused only on the substrate-binding site of the M^{pro} enzyme.[51−53] Other computational studies have looked into identifying novel pockets and investigating allostery.[54,55] However, these studies are limited in comparing dynamics with the vast crystallographic data available on ortho- and allosteric ligand-binding sites across β-CoV homologs.

In this study, we map all nonredundant ligand-binding sites reported in the PDB for β-CoV M^{pro} enzyme homologs, including SARS-CoV2, SARS-CoV, and MERS-CoV. We perform 25 μs MSM-based adaptive sampling MD simulations to study the dynamics of the binding sites. It is worth noting that we simulated the apo form of SARS-CoV2, which was generated by the removal of the ligand from the substrate-binding site in PDB 6LU7. However, this does not have any impact on our analysis as we sample all crystallographic conformations. The analysis emphasizes that even though the β-CoV M^{pro} structures are very similar, they display remarkably different structural dynamics (Figure S9). The differences in dynamics are subtle and indistinguishable using conventional methods. Therefore, we employed dynamically sensitive CVAE-based machine-learning approaches to resolve the differences between each system. MSMs were built to identify kinetically relevant metastable states, which were then used to study the spatiotemporal evolution of the ligand-binding sites. The metastable states generated from the simulations were searched for the presence of pockets and compared individually with all other experimentally derived crystal structures representing nonredundant ligand-binding sites.

The M^{pro} enzymes are homodimers, and each binding site is present as two copies, one on each protomer except for site V. The dynamical behavior of the binding sites in each protomer is stochastic and independent of the other (Figure S10). This is evident from the structural dynamics of the binding sites, which, in some metastable states, appear only in one protomer and absent in the other. Our finding is supported by a previous work on M^{pro} enzymes, where the dynamics of different protomers map on different regions of the conformational space.[51] We also identify that loops connecting different structural features are the most flexible regions of the enzyme and contribute toward local motions, while the movement between the two coaxially stacked protomers contributes to the global dynamics. The presence or absence of binding sites in each protomer is independent of the influence of the adjacent protomer except for the sites at the interface. The ligands that bind at the interface work by stabilizing the global motions that contribute toward inhibiting mechanistic functions.

To assess the possibility of the broad-spectrum inhibition of M^{pro} enzymes, we analyzed the structural and dynamic conservation of the binding sites across the three β-CoV homologs. We rationalized that an inhibitor designed to target a conserved binding site would have relatable effects across homologs. This would be advantageous for the design of therapeutics in dealing with any future viral outbreaks. We analyzed the dynamics of the ligand-binding sites by comparing the sequence and structural features between relative homologs.

SARS-CoV2 and SARS-CoV have 96% similar sequence identity. We identify that of the 12 residues (out of 306) that are different between SARS-CoV2 and SARS-CoV (Figure 1D) in each protomer, eight are associated with an experimentally identified ligand-binding site. The substitution of some of these residues have an effect on the surface charge patterns ($N180_{SARS\text{-}CoV2}$/$K180_{SARS\text{-}CoV}$ and $T35_{SARS\text{-}CoV2}$/$V35_{SARS\text{-}CoV}$), interactions ($F134_{SARS\text{-}CoV2}$/$H134_{SARS\text{-}CoV}$), and dimensions ($V202_{SARS\text{-}CoV2}$/$L202_{SARS\text{-}CoV}$), enhancing enzymatic activity via dimerization ($A285_{SARS\text{-}CoV2}$/$T285_{SARS\text{-}CoV}$ and $L286_{SARS\text{-}CoV2}$/$I286_{SARS\text{-}CoV}$) or completely blocking the space where the ligand binds ($A94_{SARS\text{-}CoV2}$/$S94_{SARS\text{-}CoV}$). One substitution ($K88_{SARS\text{-}CoV2}$/$R88_{SARS\text{-}CoV}$) has no notable effect on the binding site. Two residues ($S46_{SARS\text{-}CoV2}$/$A46_{SARS\text{-}CoV}$ and $N65_{SARS\text{-}CoV2}$/$S65_{SARS\text{-}CoV}$) are a part of potential cavities, but no ligand has been identified to bind to them yet. The remaining two residues ($V86_{SARS\text{-}CoV2}$/$L86_{SARS\text{-}CoV}$ and $S267_{SARS\text{-}CoV2}$/$A267_{SARS\text{-}CoV}$) are located in the core of the enzyme and are not solvent-accessible.

Then, we tracked the dynamic persistence of the ligand-binding sites in the MSM-derived metastable states in the three homologs and made comparisons with the representative binding sites from the crystal structures. All the identified binding sites were located on the surface of the M^{pro}. Ligand-binding sites A-L, P, R, S, and V (SARS-CoV2); A-F, H-J, L, P-S, and V (SARS-CoV); and A-C, F, K-M, P, Q, S, and V (MERS-CoV) are present in all metastable states. Site O is the only ligand-binding site that is absent in all homologs. Site O is a pseudo-binding site on a solvent-exposed loop, whose conformation once lost is never observed in the dynamics of apo M^{pro}. Sites M, N, Q, T, and U in SARS-CoV2; N, T, and U in SARS-CoV; and D, E, G, I, and T in MERS-CoV are present in some states and absent in others. Sites G, K, M, and O (SARS-CoV) as well as H, J, N, O, R, and U (MERS-CoV) are completely absent in their respective homologs. It is worth noting that there are multiple binding sites that lie adjacent to one another, for example, sites B and C; P and T; R and S. Fragments occupying these sites can be chemically linked to

https://doi.org/10.1021/acs.jcim.1c00449
J. Chem. Inf. Model. 2021, 61, 3058−3073

enhance effective binding (Figure S11). Furthermore, there are several other structural features present around the experimentally identified binding sites, which can be exploited to improve the design of inhibitors. For example, empty cavities are present adjacent to sites H, K, L, N, Q, and S (Figure S12). These cavities can be used as extensions of the existing binding sites to improve the ligand design. Moreover, there are some additional pockets that appear in all homologs at the protomer interface in domain I, which could be further exploited for the ligand design (Figure S13).

Our detailed structural dynamics analysis highlights the importance of the dynamic conservation of ligand-binding sites across β-CoV homologs. Based on these observations, we emphasize that the ligand design should be preferred on target binding sites that are not only structurally but also dynamically conserved across all β-CoV homologs.

CONCLUSIONS

The past 20 years have seen outbreaks caused by three highly pathogenic β-CoV, namely, SARS-CoV in 2002, MERS-CoV in 2013, and SARS-CoV2 in 2019.[56] The social and economic impact of the current pandemic has been exceptional. This crisis has led to an urgent requirement to develop therapeutics. Even though a number of vaccines have been approved by the Food and Drug Administration, alternative strategies targeting essential viral components are required as a backup against the emergence of lethal viral variants. One such target is the main protease that plays an indispensable role in viral replication.[18,20] Multinodal, large interdisciplinary consortiums have reported potential drug candidates.[40−42] The availability of M^{pro} X-ray structures in complex with inhibitors provides unique insights into ligand interactions. These data, in conjunction with molecular simulations, can aid in further improving the design of inhibitors, including exploring the dynamic conservation of ligand-binding sites across β-CoV homologs that are highly relevant to human diseases. Employing such a strategy is essential in preparing toward any future viral outbreaks.

EXPERIMENTAL METHODS

Ligand-Binding Site Identification. The PDB in Europe knowledge base (PDBe-KB) was searched with the keyword "3C-like proteinase" and selecting "Severe acute respiratory syndrome CoV 2 (2019-nCoV)" as the organism. The PDB codes were noted and the structural coordinates downloaded. Thorough analysis was performed by the superimposition of the structures. A binding site was defined where a chemical fragment or a compound interacted with the M^{pro} protease structure derived from crystallographic experiments. The key interacting residues in the protein were identified within a 4.0 Å cutoff distance around the ligand. This was repeated until all entries were evaluated. From this list, a nonredundant representative structure for each binding site was identified. For example, in PDB 6LU7,[22] ligand N3 interacts with residues C145, H41, G189, P168, E166, H163, and H164 in the substrate-binding site. Thus, 6LU7 was selected as the representative structure for all ligands that interacted with these residues and labeled "site A". Figures for representative structure and ligands were generated using Protein Imager[57] and OpenEye toolkits 2020.2.2 (www.eyesopen.com).

A similar protocol was applied for SARS-CoV and MERS-CoV M^{pro} structures, and nonredundant representative structures were identified after superimposition with SARS-CoV2 structures. PDB identifiers, structural analysis, and ligand interaction data are listed in the Supporting Information. The nonredundant representative ligand-binding site data have been tabulated in Tables 123.

Adaptive Sampling MD Simulations. The coordinates of the apo structures of the SARS-CoV2 (PDB 6LU7),[22] SARS-CoV (PDB 2C3S),[58] and MERS-CoV (PDB 4YLU)[59] proteases in their dimeric forms were downloaded to run MD simulations. Ligands and all crystallization agents/additives were removed from their respective binding sites. The protonation states of all titratable side chains were determined using the *ProteinPrepare* functionality, as implemented in the High throughput Molecular Dynamics (HTMD) framework.[60,61] The charges were assigned after the optimization of the hydrogen-bonding network in the protonated structure.[61] The catalytic cysteine residue was set to a reduced state. The Amber ff14SB forcefield was used to describe the protein.[62] Each system was solvated using TIP3P water in a cubic box, the edge of which was set to at least 10 Å from the closest solute atom.[63] Counterions were added to neutralize the system. The simulation protocol was identical for each system. The systems were minimized and relaxed under NPT conditions for 50 ns at 1 atm. The temperature was increased to 300 K using a time step of 4 fs, rigid bonds, cutoff of 9.0 Å, and particle mesh Ewald (PME) summations switched on for long-range electrostatics.[64] During the equilibration step, the protein's backbone was restrained by a spring constant set at 1 kcal mol^{-1} $Å^{-2}$, while the ions and the solvent were free to move. The production simulations were run in the NVT ensemble using a Langevin thermostat with a damping constant of 0.1 ps and a hydrogen mass repartitioning scheme to achieve a time step of 4 fs.[65] The final production step was run as adaptive sampling, without any restraints, as multiple iterations of short parallel simulations, as implemented in the HTMD framework.[60] Each system was run for 125 epochs (iterations), and each epoch consists of four parallel simulations of 50 ns each, which equals 25 μs of simulated time. The short simulations after each epoch are postprocessed based on the backbone dihedral angle metric. A rough Markov model is then used to decide from which part of the configuration space to respawn the following simulations in the next epoch. The visualization of the simulations was performed using the PyMOL-MDanalysis (https://github.com/bieniekmateusz/pymol-mdanalysis)[66] and the VMD package.[67]

MSMs. MSMs were constructed to provide kinetics and free energy estimates. The MSM was built using the PyEMMA v2.5.7 program.[68] It was not possible to build an MSM using just the features of the 24 dissimilar residues (12 in each protomer) between SARS-CoV2 and SARS-CoV. Therefore, all backbone dihedral angles were selected. In addition, the first χ angle (χ1) from 24 dissimilar residues was also included in MSM building. For MERS-CoV, the χ1 angles from residues at the equivalent position were also selected. Time-lagged independent component analysis (tICA) was used to reduce the dimensionality of the data.[69,70] It was possible to build models that were Markovian with a lag time of ≥10, with the lag time selected according to the convergence of the implied timescales. The dimension reduction was achieved by projecting on the three slowest tICA components. The K-means clustering algorithm was used to obtain 100 microstates. The conformational clusters were grouped together based on

https://doi.org/10.1021/acs.jcim.1c00449
J. Chem. Inf. Model. 2021, 61, 3058−3073

the kinetic similarity using the PCCA+ algorithm.[71] The PCCA+ algorithm uses the eigenvectors of the MSMs to group together clusters, which are kinetically close, resulting in a set of macrostates. The final number of metastable macrostates was selected based on the implied timescale plot. The MSMs were validated using the Chapman−Kolmogorov test implemented in PyEMMA.[68]

CVAE-Based Deep-Learning Implementation. The convolutional variational AE or CVAE was used for analysis,[43] which has been optimized for large-scale systems on the high performance computing (HPC) platform.[72] The implementation of CVAE has been previously shown to provide meaningful insights into diverse systems such as protein folding,[73] enzyme dynamics,[74,75] CoV spike protein,[76] and CoV nonstructured proteins.[77]

A CVAE consists of a variational AE along with multiple convolutional layers. Generally, an AE has an hourglass shape where high-dimensional data go into as the input, and the AE captures only the essential information required to represent the original input data. This compressed latent representation is then used to reconstruct the data back to the original format, ensuring no loss of information during the compression phase. The variational approach at the latent space is included as an additional optimization requirement. The introduction of the variational technique forces the compressed key information to normally distribute over the latent space. Convolutional layers are used instead of feed forward layers because the convolutional layers are more effective at detecting and capturing both the local and global patterns in the input data, especially where the data have multilayered structures like complex proteins, as presented here. The complete CVAE structure is shown in Figure S3A with different steps that are performed from raw simulation data to the resolution of β-CoV M^{pro} solely based on their local and global conformational dynamics.

The distance matrix of the 24 × 24 dissimilar Cα atoms was used as the input for the CVAE architecture. Using the Horovod library, the data parallel model was trained on the Summit supercomputer. Each CVAE was trained for a fixed number of epochs based on the convergence of loss and variance-bias tradeoff. Each training utilized up to 16 Summit nodes (96V100 GPUs), with the effective batch size being the sum of every individual training instance. Therefore, the individual batch size was selected to be relatively small to avoid the generalization gap for large-batch training. The data set was divided into training/validation (80:20% of the simulation trajectories) and randomly shuffled. To search for the optimal clustering and reconstruction quality of the CVAE, the training procedure was repeated for various latent dimension sizes and to identify the best model for the data set (Figure S3B). The loss over the epochs is as expected (i.e., without over fitting or any other unusual behavior) and shown in Figure S3C. Finally, the original input data were compared with the predicted (i.e., decompressed) data to ensure no loss of information during the compression process through the latent space (Figure S3D).

Dynamic Pocket Tracking. Pocketron was used to detect small molecule binding sites using default values.[78] The metastable states were screened for pockets, which were classified as open if they could accommodate at least five water molecules (coarse equivalent of a small fragment). Each representative binding pocket, identified from the crystal structures, was compared by superimposition with the metastable state from each system.

Analysis of Pairwise Correlated Positions in Evolution. Pairwise evolutionary constraints were estimated from a multiple sequence alignment (MSA). The FASTA sequence from the SARS-CoV2 M^{pro} (PDB 6LU7) was selected as the reference, and the MSA was built using hhsuite3.[79] Pairwise correlations were calculated using the ccmpred package[80] as per the parameters described by Akere et al.[74] Raw correlation scores (C_i) were then scaled as reported by Kamisetty et al.[81] For all 22 pockets (see Table 1), the scaled pairwise correlation matrix was used to estimate the evolutionary conservation score (E_a) of each pocket (eq 1), where N is the number of residues in the pocket.

$$E_A = \frac{1}{N}\sum_{i=1}^{N}\sum_{j>i}^{N} C_{ij}/N \quad (1)$$

The score estimates the evolutionary constraints on the pocket as an average of the pairwise correlation in the pocket. For reference, scores were compared with the median and standard deviation of C_i for all surface residue pairs (Figure S10). Surface residues were defined as having >50% relative accessible surface area.[82,83]

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jcim.1c00449.

Mapping the binding sites; CVAE-based deep learning; interactions between the SARS-CoV2 M^{pro} in the pseudo-ligand-binding sites; conformational drift in M^{pro} enzymes; CVAE-based deep-learning implementation; conventional RMS fluctuation analysis; MSM of SARS-CoV2, SARS-CoV, and MERS-CoV M^{pro} enzymes; evolution conservation score for each pocket; superimposition of SARS-CoV2 and SARS-CoV M^{pro} structures; RMS fluctuation plots after fraction alignment; adjacent binding sites in SARS-CoV2; multiple ligand binding in sites; and potential binding sites (PDF)

Details of the ligand-binding sites (PDF)

■ AUTHOR INFORMATION

Corresponding Author

Shozeb Haider − *UCL School of Pharmacy, London WC1N 1AX, U.K.*; orcid.org/0000-0003-2650-2925; Email: Shozeb.haider@ucl.ac.uk

Authors

Eunice Cho − *UCL School of Pharmacy, London WC1N 1AX, U.K.*

Margarida Rosa − *UCL School of Pharmacy, London WC1N 1AX, U.K.*

Ruhi Anjum − *Department of Biochemistry, Aligarh Muslim University, Aligarh, Uttar Pradesh 202002, India*

Saman Mehmood − *Department of Zoology, Aligarh Muslim University, Aligarh, Uttar Pradesh 202002, India*

Mariya Soban − *Department of Biochemistry, Aligarh Muslim University, Aligarh, Uttar Pradesh 202002, India*

Moniza Mujtaba − *Herricks High School, New Hyde Park, New York 11040, United States*

Khair Bux − *Third World Center for Science and Technology, H.E.J. Research Institute of Chemistry, International Centre*

of Chemical and Biological Sciences, University of Karachi, Karachi 75270, Pakistan

Syed T. Moin − *Third World Center for Science and Technology, H.E.J. Research Institute of Chemistry, International Centre of Chemical and Biological Sciences, University of Karachi, Karachi 75270, Pakistan;* orcid.org/0000-0002-2868-7663

Mohammad Tanweer − *UCL School of Pharmacy, London WC1N 1AX, U.K.*

Sarath Dantu − *Department of Computer Science, Brunel University, Uxbridge UB8 3PH, U.K.;* orcid.org/0000-0003-2019-5311

Alessandro Pandini − *Department of Computer Science, Brunel University, Uxbridge UB8 3PH, U.K.;* orcid.org/0000-0002-4158-233X

Junqi Yin − *Center for Computational Sciences, Oak Ridge National Laboratory, Oak Ridge, Tennessee 37830, United States*

Heng Ma − *Data Science and Learning Division, Argonne National Laboratory, Lemont, Illinois 60439, United States*

Arvind Ramanathan − *Data Science and Learning Division, Argonne National Laboratory, Lemont, Illinois 60439, United States; Consortium for Advanced Science and Engineering, University of Chicago, Chicago, Illinois 60637, United States*

Barira Islam − *Department of Bioscience, University of Huddersfield, Huddersfield HD1 3DH, U.K.;* orcid.org/0000-0001-5882-6903

Antonia S. J. S. Mey − *EaStCHEM School of Chemistry, University of Edinburgh, Edinburgh EH9 3FJ, U.K.;* orcid.org/0000-0001-7512-5252

Debsindhu Bhowmik − *Computer Sciences and Engineering Division, Oak Ridge National Laboratory, Oak Ridge, Tennessee 37830, United States;* orcid.org/0000-0001-7770-9091

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jcim.1c00449

Author Contributions

Data mining and collation: E.C., M.R., R.A., S.M., M.S., M.M., K.B., and B.I.; Binding site analysis: E.C., K.B., B.I., and S.H.; Coevolution analysis: S.D. and A.P.; Deep learning: H.M., A.R., D.B., J.Y., M.T., and S.H.; Simulations: S.H., H.M., A.R.; M.S.M., S.H., and A.M.; Manuscript writing: E.C., A.R., D.B., and S.H.; Other inputs: all coauthors.

Notes

This manuscript has been authored by UT-Battelle, LLC under Contract No. DE-AC05-00OR22725 with the U.S. Department of Energy. The United States Government retains and the publisher, by accepting the article for publication, acknowledges that the United States Government retains a nonexclusive, paid-up, irrevocable, worldwide license to publish or reproduce the published form of this manuscript, or allow others to do so, for United States Government purposes. The Department of Energy will provide public access to these results of federally sponsored research in accordance with the DOE Public Access Plan (http://energy.gov/downloads/doe-public-accessplan).

The authors declare no competing financial interest.

The trajectories of M^{pro} simulations and models of the metastable states can be downloaded from 10.5281/zenodo.4782284

■ ACKNOWLEDGMENTS

The authors would like to thank the UK High-End Computing Consortium for Biomolecular Simulation, HECBioSim (http://hecbiosim.ac.uk) for time to run simulations on ARCHER. B.I. would like to acknowledge the COVID-19 pump-priming grant from the University of Huddersfield for funding computing resources for analysis. S.H. would like to thank Prof Frank Kozielski for insightful discussions on the manuscript. This material is based upon the work supported by the U.S. Department of Energy, Office of Science, through the Advanced Scientific Computing Research (ASCR), under contract number DEAC05-00OR22725 and the Exascale Computing Project (ECP) (17-SC-20-SC). This work was performed at the Oak Ridge Leadership Computing Facility (OLCF) of the Oak Ridge National Laboratory (ORNL) and used the Extreme Science and Engineering Discovery Environment (XSEDE)[84] COVID-19HPC Consortium at the IBM AC922 Summit supercomputer of the OLCF at ORNL through allocation TG-ASC200020. D.B. would like to thank ASCR and ECP for assistance in the implementation of deep-learning, data processing, and data analysis algorithms and would like to thank members of the DOE National Virtual Biotechnology Laboratory (NVBL) projects for insightful discussions on the design of and results from the calculations described in this manuscript.

■ REFERENCES

(1) Weiss, S. R. Forty Years with Coronaviruses. *J. Exp. Med.* **2020**, *217*, No. e20200537.

(2) Weiss, S. R.; Leibowitz, J. L. Coronavirus Pathogenesis. *Adv. Virus Res.* **2011**, *81*, 85−164.

(3) Woo, P. C. Y.; Lau, S. K. P.; Lam, C. S. F.; Lau, C. C. Y.; Tsang, A. K. L.; Lau, J. H. N.; Bai, R.; Teng, J. L. L.; Tsang, C. C. C.; Wang, M.; Zheng, B.-J.; Chan, K.-H.; Yuen, K.-Y. Discovery of Seven Novel Mammalian and Avian Coronaviruses in the Genus Deltacoronavirus Supports Bat Coronaviruses as the Gene Source of Alphacoronavirus and Betacoronavirus and Avian Coronaviruses as the Gene Source of Gammacoronavirus and Deltacoronavirus. *J. Virol.* **2012**, *86*, 3995−4008.

(4) Hamre, D.; Procknow, J. J. A New Virus Isolated from the Human Respiratory Tract. *Proc. Soc. Exp. Biol. Med.* **1966**, *121*, 190−193.

(5) McIntosh, K.; Dees, J. H.; Becker, W. B.; Kapikian, A. Z.; Chanock, R. M. Recovery in Tracheal Organ Cultures of Novel Viruses from Patients with Respiratory Disease. *Proc. Natl. Acad. Sci. U. S. A.* **1967**, *57*, 933−940.

(6) Vabret, A.; Dina, J.; Gouarin, S.; Petitjean, J.; Corbet, S.; Freymuth, F. Detection of the New Human Coronavirus HKU1: A Report of 6 Cases. *Clin. Infect. Dis.* **2006**, *42*, 634−639.

(7) Geller, C.; Varbanov, M.; Duval, R. E. Human Coronaviruses: Insights into Environmental Resistance and Its Influence on the Development of New Antiseptic Strategies. *Viruses* **2012**, *4*, 3044−3068.

(8) Fehr, A. R.; Perlman, S. Coronaviruses: An Overview of Their Replication and Pathogenesis. *Methods Mol. Biol.* **2015**, *1282*, 1−23.

(9) de Wit, E.; van Doremalen, N.; Falzarano, D.; Munster, V. J. SARS and MERS: Recent Insights into Emerging Coronaviruses. *Nat. Rev. Microbiol.* **2016**, *14*, 523−534.

(10) Greenberg, S. B. Update on Human Rhinovirus and Coronavirus Infections. *Semin. Respir. Crit. Care. Med.* **2016**, *37*, 555−571.

(11) Vos, L. M.; Bruyndonckx, R.; Zuithoff, N. P. A.; Little, P.; Oosterheert, J. J.; Broekhuizen, B. D. L.; Lammens, C.; Loens, K.; Viveen, M.; Butler, C. C.; Crook, D.; Zlateva, K.; Goossens, H.; Claas, E. C. J.; Ieven, M.; Van Loon, A. M.; Verheij, T. J. M.; Coenjaerts, F. E. J. Lower Respiratory Tract Infection in the Community:

Associations between Viral Aetiology and Illness Course. *Clin. Microbiol. Infect.* **2021**, *27*, 96−104.
(12) Petersen, E.; Koopmans, M.; Go, U.; Hamer, D. H.; Petrosillo, N.; Castelli, F.; Storgaard, M.; Khalili, S. A.; Simonsen, L. Comparing SARS-CoV-2 with SARS-CoV and Influenza Pandemics. *Lancet Infect. Dis.* **2020**, *20*, e238−e244.
(13) Weiss, S. R.; Navas-Martin, S. Coronavirus Pathogenesis and the Emerging Pathogen Severe Acute Respiratory Syndrome Coronavirus. *Microbiol. Mol. Biol. Rev.* **2005**, *69*, 635−664.
(14) Masters, P. S. The Molecular Biology of Coronaviruses. *Adv. Virus Res.* **2006**, *66*, 193−292.
(15) V'kovski, P.; Kratzel, A.; Steiner, S.; Stalder, H.; Thiel, V. Coronavirus Biology and Replication: Implications for SARS-CoV-2. *Nat. Rev. Microbiol.* **2021**, *19*, 155−170.
(16) Finkel, Y.; Mizrahi, O.; Nachshon, A.; Weingarten-Gabbay, S.; Morgenstern, D.; Yahalom-Ronen, Y.; Tamir, H.; Achdout, H.; Stein, D.; Israeli, O.; Beth-Din, A.; Melamed, S.; Weiss, S.; Israely, T.; Paran, N.; Schwartz, M.; Stern-Ginossar, N. The Coding Capacity of SARS-CoV-2. *Nature* **2021**, *589*, 125−130.
(17) Michel, C. J.; Mayer, C.; Poch, O.; Thompson, J. D. Characterization of Accessory Genes in Coronavirus Genomes. *Virol. J.* **2020**, *17*, 131.
(18) Ullrich, S.; Nitsche, C. The SARS-CoV-2 Main Protease as Drug Target. *Bioorg. Med. Chem. Lett.* **2020**, *30*, No. 127377.
(19) Hilgenfeld, R. From SARS to MERS: Crystallographic Studies on Coronaviral Proteases Enable Antiviral Drug Design. *FEBS J.* **2014**, *281*, 4085−4096.
(20) Rut, W.; Groborz, K.; Zhang, L.; Sun, X.; Zmudzinski, M.; Pawlik, B.; Wang, X.; Jochmans, D.; Neyts, J.; Młynarski, W.; Hilgenfeld, R.; Drag, M. SARS-CoV-2 M pro Inhibitors and Activity-Based Probes for Patient-Sample Imaging. *Nat. Chem. Biol.* **2021**, *17*, 222−228.
(21) Zhang, L.; Lin, D.; Sun, X.; Curth, U.; Drosten, C.; Sauerhering, L.; Becker, S.; Rox, K.; Hilgenfeld, R. Crystal Structure of SARS-CoV-2 Main Protease Provides a Basis for Design of Improved α-Ketoamide Inhibitors. *Science* **2020**, *368*, 409−412.
(22) 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 M pro from SARS-CoV-2 and Discovery of Its Inhibitors. *Nature* **2020**, *582*, 289−293.
(23) Paasche, A.; Zipper, A.; Schäfer, S.; Ziebuhr, J.; Schirmeister, T.; Engels, B. Evidence for Substrate Binding-Induced Zwitterion Formation in the Catalytic Cys-His Dyad of the SARS-CoV Main Protease. *Biochemistry* **2014**, *53*, 5930−5946.
(24) Świderek, K.; Moliner, V. Revealing the Molecular Mechanisms of Proteolysis of SARS-CoV-2 Mpro by QM/MM Computational Methods. *Chem. Sci.* **2020**, *11*, 10626−10630.
(25) Pathak, N.; Chen, Y.-T.; Hsu, Y.-C.; Hsu, N.-Y.; Kuo, C.-J.; Tsai, H. P.; Kang, J.-J.; Huang, C.-H.; Chang, S.-Y.; Chang, Y.-H.; Liang, P.-H.; Yang, J.-M. Uncovering Flexible Active Site Conformations of SARS-CoV-2 3CL Proteases through Protease Pharmacophore Clusters and COVID-19 Drug Repurposing. *ACS Nano* **2021**, *15*, 857−872.
(26) Hsu, W.-C.; Chang, H.-C.; Chou, C.-Y.; Tsai, P.-J.; Lin, P.-I.; Chang, G.-G. Critical Assessment of Important Regions in the Subunit Association and Catalytic Action of the Severe Acute Respiratory Syndrome Coronavirus Main Protease. *J. Biol. Chem.* **2005**, *280*, 22741−22748.
(27) Xia, B.; Kang, X. Activation and Maturation of SARS-CoV Main Protease. *Protein Cell* **2011**, *2*, 282−290.
(28) 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.
(29) Jin, Z.; Zhao, Y.; Sun, Y.; Zhang, B.; Wang, H.; Wu, Y.; Zhu, Y.; Zhu, C.; Hu, T.; Du, X.; Duan, Y.; Yu, J.; Yang, X.; Yang, X.; Yang, K.; Liu, X.; Guddat, L. W.; Xiao, G.; Zhang, L.; Yang, H.; Rao, Z. Structural Basis for the Inhibition of SARS-CoV-2 Main Protease by Antineoplastic Drug Carmofur. *Nat. Struct. Mol. Biol.* **2020**, *27*, 529−532.
(30) Arafet, K.; Serrano-Aparicio, N.; Lodola, A.; Mulholland, A. J.; González, F. V.; Świderek, K.; Moliner, V. Mechanism of Inhibition of SARS-CoV-2 Mpro by N3 Peptidyl Michael Acceptor Explained by QM/MM Simulations and Design of New Derivatives with Tunable Chemical Reactivity. *Chem. Sci.* **2021**, *12*, 1433−1444.
(31) Chen, Y. W.; Yiu, C.-P. B.; Wong, K.-Y. Prediction of the SARS-CoV-2 (2019-NCoV) 3C-like Protease (3CLpro) Structure: Virtual Screening Reveals Velpatasvir, Ledipasvir, and Other Drug Repurposing Candidates. *F1000Res* **2020**, *9*, 129.
(32) Elmezayen, A. D.; Al-Obaidi, A.; Şahin, A. T.; Yelekçi, K. Drug Repurposing for Coronavirus (COVID-19): In Silico Screening of Known Drugs against Coronavirus 3CL Hydrolase and Protease Enzymes. *J. Biomol. Struct. Dyn.* **2021**, *39*, 2980−2992.
(33) Ghahremanpour, M. M.; Tirado-Rives, J.; Deshmukh, M.; Ippolito, J. A.; Zhang, C.-H.; Cabeza de Vaca, I.; Liosi, M.-E.; Anderson, K. S.; Jorgensen, W. L. Identification of 14 Known Drugs as Inhibitors of the Main Protease of SARS-CoV-2. *ACS Med. Chem. Lett.* **2020**, *11*, 2526−2533.
(34) Hofmarcher, M.; Mayr, A.; Rumetshofer, E.; Ruch, P.; Renz, P.; Schimunek, J.; Seidl, P.; Vall, A.; Widrich, M.; Hochreiter, S.; Klambauer, G. *Large-Scale Ligand-Based Virtual Screening for SARS-CoV-2 Inhibitors Using Deep Neural Networks; SSRN Scholarly Paper ID 3561442*; Social Science Research Network: Rochester, NY, 2020.
(35) Kandeel, M.; Al-Nazawi, M. Virtual Screening and Repurposing of FDA Approved Drugs against COVID-19 Main Protease. *Life Sci.* **2020**, *251*, No. 117627.
(36) Ton, A.-T.; Gentile, F.; Hsing, M.; Ban, F.; Cherkasov, A. Rapid Identification of Potential Inhibitors of SARS-CoV-2 Main Protease by Deep Docking of 1.3 Billion Compounds. *Mol. Inform.* **2020**, *39*, No. 2000028.
(37) Cannalire, R.; Cerchia, C.; Beccari, A. R.; Di Leva, F. S.; Summa, V. *Targeting SARS-CoV-2 Proteases and Polymerase for COVID-19 Treatment: State of the Art and Future Opportunities*. J. Med. Chem. 2020, DOI: 10.1021/acs.jmedchem.0c01140.
(38) Cui, W.; Yang, K.; Yang, H. Recent Progress in the Drug Development Targeting SARS-CoV-2 Main Protease as Treatment for COVID-19. *Front. Mol. Biosci.* **2020**, *7*, No. 616341.
(39) Banerjee, S.; Pieper, U.; Kapadia, G.; Pannell, L. K.; Herzberg, O. Role of the Ω-Loop in the Activity, Substrate Specificity, and Structure of Class A β-Lactamase. *Biochemistry* **1998**, *37*, 3286−3296.
(40) Consortium, T. C. M.; Achdout, H.; Aimon, A.; Bar-David, E.; Barr, H.; Ben-Shmuel, A.; Bennett, J.; Bobby, M. L.; Brun, J.; Sarma, B.; Calmiano, M.; Carbery, A.; Cattermole, E.; Chodera, J. D.; Clyde, A.; Coffland, J. E.; Cohen, G.; Cole, J.; Contini, A.; Cox, L.; Cvitkovic, M.; Dias, A.; Douangamath, A.; Duberstein, S.; Dudgeon, T.; Dunnett, L.; Eastman, P. K.; Erez, N.; Fairhead, M.; Fearon, D.; Fedorov, O.; Ferla, M.; Foster, H.; Foster, R.; Gabizon, R.; Gehrtz, P.; Gileadi, C.; Giroud, C.; Glass, W. G.; Glen, R.; Glinert, I.; Gorichko, M.; Gorrie-Stone, T.; Griffen, E. J.; Heer, J.; Hill, M.; Horrell, S.; Hurley, M. F. D.; Israely, T.; Jajack, A.; Jnoff, E.; John, T.; Kantsadi, A. L.; Kenny, P. W.; Kiappes, J. L.; Koekemoer, L.; Kovar, B.; Krojer, T.; Lee, A. A.; Lefker, B. A.; Levy, H.; London, N.; Lukacik, P.; Macdonald, H. B.; MacLean, B.; Malla, T. R.; Matviiuk, T.; McCorkindale, W.; Melamed, S.; Michurin, O.; Mikolajek, H.; Morris, A.; Morris, G. M.; Morwitzer, M. J.; Moustakas, D.; Neto, J. B.; Oleinikovas, V.; Overheul, G. J.; Owen, D.; Pai, R.; Pan, J.; Paran, N.; Perry, B.; Pingle, M.; Pinjari, J.; Politi, B.; Powell, A.; Psenak, V.; Puni, R.; Rangel, V. L.; Reddi, R. N.; Reid, S. P.; Resnick, E.; Robinson, M. C.; Robinson, R. P.; Rufa, D.; Schofield, C.; Shaikh, A.; Shi, J.; Shurrush, K.; Sittner, A.; Skyner, R.; Smalley, A.; Smilova, M. D.; Spencer, J.; Strain-Damerell, C.; Swamy, V.; Tamir, H.; Tennant, R.; Thompson, A.; Thompson, W.; Tomasio, S.; Tumber, A.; Vakonakis, I.; Rij, R. P. van; Varghese, F. S.; Vaschetto, M.; Vitner,

E. B.; Voelz, V.; Delft, A. von; Delft, F. von; Walsh, M.; Ward, W.; Weatherall, C.; Weiss, S.; Wild, C. F.; Wittmann, M.; Wright, N.; Yahalom-Ronen, Y.; Zaidmann, D.; Zidane, H.; Zitzmann, N. *COVID Moonshot: Open Science Discovery of SARS-CoV-2 Main Protease Inhibitors by Combining Crowdsourcing, High-Throughput Experiments, Computational Simulations, and Machine Learning*. bioRxiv 2020, DOI: 10.1101/2020.10.29.339317.

(41) Douangamath, A.; Fearon, D.; Gehrtz, P.; Krojer, T.; Lukacik, P.; Owen, C. D.; Resnick, E.; Strain-Damerell, C.; Aimon, A.; Ábrányi-Balogh, P.; Brandão-Neto, J.; Carbery, A.; Davison, G.; Dias, A.; Downes, T. D.; Dunnett, L.; Fairhead, M.; Firth, J. D.; Jones, S. P.; Keeley, A.; Keserü, G. M.; Klein, H. F.; Martin, M. P.; Noble, M. E. M.; O'Brien, P.; Powell, A.; Reddi, R. N.; Skyner, R.; Snee, M.; Waring, M. J.; Wild, C.; London, N.; von Delft, F.; Walsh, M. A. Crystallographic and Electrophilic Fragment Screening of the SARS-CoV-2 Main Protease. *Nat. Commun.* **2020**, *11*, 5047.

(42) Günther, S.; Reinke, P. Y. A.; Fernández-García, Y.; Lieske, J.; Lane, T. J.; Ginn, H. M.; Koua, F. H. M.; Ehrt, C.; Ewert, W.; Oberthuer, D.; Yefanov, O.; Meier, S.; Lorenzen, K.; Krichel, B.; Kopicki, J.-D.; Gelisio, L.; Brehm, W.; Dunkel, I.; Seychell, B.; Gieseler, H.; Norton-Baker, B.; Escudero-Pérez, B.; Domaracky, M.; Saouane, S.; Tolstikova, A.; White, T. A.; Hänle, A.; Groessler, M.; Fleckenstein, H.; Trost, F.; Galchenkova, M.; Gevorkov, Y.; Li, C.; Awel, S.; Peck, A.; Barthelmess, M.; Schlünzen, F.; Lourdu Xavier, P.; Werner, N.; Andaleeb, H.; Ullah, N.; Falke, S.; Srinivasan, V.; França, B. A.; Schwinzer, M.; Brognaro, H.; Rogers, C.; Melo, D.; Zaitseva-Doyle, J. J.; Knoska, J.; Peña-Murillo, G. E.; Mashhour, A. R.; Hennicke, V.; Fischer, P.; Hakanpää, J.; Meyer, J.; Gribbon, P.; Ellinger, B.; Kuzikov, M.; Wolf, M.; Beccari, A. R.; Bourenkov, G.; von Stetten, D.; Pompidor, G.; Bento, I.; Panneerselvam, S.; Karpics, I.; Schneider, T. R.; Garcia-Alai, M. M.; Niebling, S.; Günther, C.; Schmidt, C.; Schubert, R.; Han, H.; Boger, J.; Monteiro, D. C. F.; Zhang, L.; Sun, X.; Pletzer-Zelgert, J.; Wollenhaupt, J.; Feiler, C. G.; Weiss, M. S.; Schulz, E.-C.; Mehrabi, P.; Karničar, K.; Usenik, A.; Loboda, J.; Tidow, H.; Chari, A.; Hilgenfeld, R.; Uetrecht, C.; Cox, R.; Zaliani, A.; Beck, T.; Rarey, M.; Günther, S.; Turk, D.; Hinrichs, W.; Chapman, H. N.; Pearson, A. R.; Betzel, C.; Meents, A. X-Ray Screening Identifies Active Site and Allosteric Inhibitors of SARS-CoV-2 Main Protease. *Science* **2021**, *372*, 642−646.

(43) Bhowmik, D.; Gao, S.; Young, M. T.; Ramanathan, A. Deep Clustering of Protein Folding Simulations. *BMC Bioinformatics* **2018**, *19*, 484.

(44) Shukla, D.; Hernández, C. X.; Weber, J. K.; Pande, V. S. Markov State Models Provide Insights into Dynamic Modulation of Protein Function. *Acc. Chem. Res.* **2015**, *48*, 414−422.

(45) Juárez-Jiménez, J.; Gupta, A. A.; Karunanithy, G.; Mey, A. S. J. S.; Georgiou, C.; Ioannidis, H.; De Simone, A.; Barlow, P. N.; Hulme, A. N.; Walkinshaw, M. D.; Baldwin, A. J.; Michel, J. Dynamic Design: Manipulation of Millisecond Timescale Motions on the Energy Landscape of Cyclophilin A. *Chem. Sci.* **2020**, *11*, 2670−2680.

(46) Buch, I.; Giorgino, T.; De Fabritiis, G. Complete Reconstruction of an Enzyme-Inhibitor Binding Process by Molecular Dynamics Simulations. *Proc. Natl. Acad. Sci. U. S. A.* **2011**, *108*, 10184−10189.

(47) Plattner, N.; Noé, F. Protein Conformational Plasticity and Complex Ligand-Binding Kinetics Explored by Atomistic Simulations and Markov Models. *Nat. Commun.* **2015**, *6*, 7653.

(48) Noé, F.; Rosta, E. Markov Models of Molecular Kinetics. *J. Chem. Phys.* **2019**, *151*, 190401.

(49) Suárez, E.; Wiewiora, R. P.; Wehmeyer, C.; Noé, F.; Chodera, J. D.; Zuckerman, D. M. What Markov State Models Can and Cannot Do: Correlation versus Path-Based Observables in Protein Folding Models. *Biophysics* **2021**, *17*, 3119−3133.

(50) Shi, J.; Song, J. The Catalysis of the SARS 3C-like Protease Is under Extensive Regulation by Its Extra Domain. *FEBS J.* **2006**, *273*, 1035−1045.

(51) Grottesi, A.; Bešker, N.; Emerson, A.; Manelfi, C.; Beccari, A. R.; Frigerio, F.; Lindahl, E.; Cerchia, C.; Talarico, C. Computational Studies of SARS-CoV-2 3CLpro: Insights from MD Simulations. *Int. J. Mol. Sci.* **2020**, *21*, 5346−5360.

(52) Suárez, D.; Díaz, N. SARS-CoV-2 Main Protease: A Molecular Dynamics Study. *J. Chem. Inf. Model.* **2020**, *60*, 5815−5831.

(53) Wan, H.; Aravamuthan, V.; Pearlstein, R. A. Probing the Dynamic Structure-Function and Structure-Free Energy Relationships of the Coronavirus Main Protease with Biodynamics Theory. *ACS Pharmacol. Transl. Sci.* **2020**, *3*, 1111−1143.

(54) Sztain, T.; Amaro, R.; McCammon, J. A. *Elucidation of Cryptic and Allosteric Pockets within the SARS-CoV-2 Protease*. Biophysics, 2020, DOI: 10.1101/2020.07.23.218784.

(55) Dubanevics, I.; McLeish, T. C. B. Computational Analysis of Dynamic Allostery and Control in the SARS-CoV-2 Main Protease. *J. R. Soc., Interface* **2021**, *18*, No. 20200591.

(56) Zhu, Z.; Lian, X.; Su, X.; Wu, W.; Marraro, G. A.; Zeng, Y. From SARS and MERS to COVID-19: A Brief Summary and Comparison of Severe Acute Respiratory Infections Caused by Three Highly Pathogenic Human Coronaviruses. *Respir. Res.* **2020**, *21*, 224.

(57) Tomasello, G.; Armenia, I.; Molla, G. The Protein Imager: A Full-Featured Online Molecular Viewer Interface with Server-Side HQ-Rendering Capabilities. *Bioinformatics* **2020**, *36*, 2909−2911.

(58) Xu, T.; Ooi, A.; Lee, H. C.; Wilmouth, R.; Liu, D. X.; Lescar, J. Structure of the SARS Coronavirus Main Proteinase as an Active C_2 Crystallographic Dimer. *Acta Crystallogr. F Struct. Biol. Cryst. Commun.* **2005**, *61*, 964−966.

(59) Tomar, S.; Johnston, M. L.; St. John, S. E.; Osswald, H. L.; Nyalapatla, P. R.; Paul, L. N.; Ghosh, A. K.; Denison, M. R.; Mesecar, A. D. Ligand-Induced Dimerization of Middle East Respiratory Syndrome (MERS) Coronavirus Nsp5 Protease (3CLpro). *J. Biol. Chem.* **2015**, *290*, 19403−19422.

(60) Doerr, S.; Harvey, M. J.; Noé, F.; De Fabritiis, G. HTMD: High-Throughput Molecular Dynamics for Molecular Discovery. *J. Chem. Theory Comput.* **2016**, *12*, 1845−1852.

(61) Martínez-Rosell, G.; Giorgino, T.; De Fabritiis, G. PlayMolecule ProteinPrepare: A Web Application for Protein Preparation for Molecular Dynamics Simulations. *J. Chem. Inf. Model.* **2017**, *57*, 1511−1516.

(62) Maier, J. A.; Martinez, C.; Kasavajhala, K.; Wickstrom, L.; Hauser, K. E.; Simmerling, C. Ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from Ff99SB. *J. Chem. Theory Comput.* **2015**, *11*, 3696−3713.

(63) Mark, P.; Nilsson, L. Structure and Dynamics of the TIP3P, SPC, and SPC/E Water Models at 298 K. *J. Phys. Chem. A* **2001**, *105*, 9954−9960.

(64) Essmann, U.; Perera, L.; Berkowitz, M. L.; Darden, T.; Lee, H.; Pedersen, L. G. A Smooth Particle Mesh Ewald Method. *J. Chem. Phys.* **1995**, *103*, 8577−8593.

(65) Feenstra, K. A.; Hess, B.; Berendsen, H. J. C. Improving Efficiency of Large Time−scale Molecular Dynamics Simulations of Hydrogen−rich Systems. *J. Comput. Chem.* **1999**, *20*, 786−798.

(66) Bieniek, M. *Fibronectin Iii_{9_10}; Adsorption to Self-Assembled Monolayers and Interdomain Orientation in the Context of Material-Driven Fibrillogenesis Studied with Molecular Dynamics Simulations*. 2020.

(67) Humphrey, W.; Dalke, A.; Schulten, K. VMD: Visual Molecular Dynamics. *J. Mol. Graph.* **1996**, *14*, 33−38.

(68) Scherer, M. K.; Trendelkamp-Schroer, B.; Paul, F.; Pérez-Hernández, G.; Hoffmann, M.; Plattner, N.; Wehmeyer, C.; Prinz, J.-H.; Noé, F. PyEMMA 2: A Software Package for Estimation, Validation, and Analysis of Markov Models. *J. Chem. Theory Comput.* **2015**, *11*, 5525−5542.

(69) Pérez-Hernández, G.; Paul, F.; Giorgino, T.; De Fabritiis, G.; Noé, F. Identification of Slow Molecular Order Parameters for Markov Model Construction. *J. Chem. Phys.* **2013**, *139*, No. 015102.

(70) Schwantes, C. R.; Pande, V. S. Improvements in Markov State Model Construction Reveal Many Non-Native Interactions in the Folding of NTL9. *J. Chem. Theory Comput.* **2013**, *9*, 2000−2009.

(71) Deuflhard, P.; Weber, M. Robust Perron Cluster Analysis in Conformation Dynamics. *Linear Algebra Appl.* **2005**, *398*, 161−184.

(72) Yoginath, S.; Alam, M.; Ramanathan, A.; Bhowmik, D.; Laanait, N.; Perumalla, K. S. Towards Native Execution of Deep Learning on a Leadership-Class HPC System. In *2019 IEEE International Parallel and Distributed Processing Symposium Workshops (IPDPSW)*; IEEE: Rio de Janeiro, Brazil, 2019; 941−950.
(73) Lee, H.; Turilli, M.; Jha, S.; Bhowmik, D.; Ma, H.; Ramanathan, A. DeepDriveMD: Deep-Learning Driven Adaptive Molecular Simulations for Protein Folding. In *2019 IEEE/ACM Third Workshop on Deep Learning on Supercomputers (DLS)*; IEEE: Denver, CO, USA, 2019; 12−19.
(74) Akere, A.; Chen, S. H.; Liu, X.; Chen, Y.; Dantu, S. C.; Pandini, A.; Bhowmik, D.; Haider, S. Structure-Based Enzyme Engineering Improves Donor-Substrate Recognition of Arabidopsis Thaliana Glycosyltransferases. *Biochem. J.* **2020**, *477*, 2791−2805.
(75) Romero, R.; Ramanathan, A.; Yuen, T.; Bhowmik, D.; Mathew, M.; Munshi, L. B.; Javaid, S.; Bloch, M.; Lizneva, D.; Rahimova, A.; Khan, A.; Taneja, C.; Kim, S.-M.; Sun, L.; New, M. I.; Haider, S.; Zaidi, M. Mechanism of Glucocerebrosidase Activation and Dysfunction in Gaucher Disease Unraveled by Molecular Dynamics and Deep Learning. *Proc. Natl. Acad. Sci. U. S. A.* **2019**, *116*, 5086−5095.
(76) Chen, S. H.; Young, M. T.; Gounley, J.; Stanley, C.; Bhowmik, D. *Distinct Structural Flexibility within SARS-CoV-2 Spike Protein Reveals Potential Therapeutic Targets*; Biophysics, 2020.
(77) Acharya, A.; Agarwal, R.; Baker, M. B.; Baudry, J.; Bhowmik, D.; Boehm, S.; Byler, K. G.; Chen, S. Y.; Coates, L.; Cooper, C. J.; Demerdash, O.; Daidone, I.; Eblen, J. D.; Ellingson, S.; Forli, S.; Glaser, J.; Gumbart, J. C.; Gunnels, J.; Hernandez, O.; Irle, S.; Kneller, D. W.; Kovalevsky, A.; Larkin, J.; Lawrence, T. J.; LeGrand, S.; Liu, S.-H.; Mitchell, J. C.; Park, G.; Parks, J. M.; Pavlova, A.; Petridis, L.; Poole, D.; Pouchard, L.; Ramanathan, A.; Rogers, D. M.; Santos-Martins, D.; Scheinberg, A.; Sedova, A.; Shen, Y.; Smith, J. C.; Smith, M. D.; Soto, C.; Tsaris, A.; Thavappiragasam, M.; Tillack, A. F.; Vermaas, J. V.; Vuong, V. Q.; Yin, J.; Yoo, S.; Zahran, M.; Zanetti-Polzi, L. Supercomputer-Based Ensemble Docking Drug Discovery Pipeline with Application to Covid-19. *J. Chem. Inf. Model.* **2020**, *60*, 5832−5852.
(78) Decherchi, S.; Bottegoni, G.; Spitaleri, A.; Rocchia, W.; Cavalli, A. BiKi Life Sciences: A New Suite for Molecular Dynamics and Related Methods in Drug Discovery. *J. Chem. Inf. Model.* **2018**, *58*, 219−224.
(79) Steinegger, M.; Meier, M.; Mirdita, M.; Vöhringer, H.; Haunsberger, S. J.; Söding, J. HH-Suite3 for Fast Remote Homology Detection and Deep Protein Annotation. *BMC Bioinformatics* **2019**, *20*, 473.
(80) Balakrishnan, S.; Kamisetty, H.; Carbonell, J. G.; Lee, S.-I.; Langmead, C. J. Learning Generative Models for Protein Fold Families: Generative Models for Protein Fold Families. *Proteins* **2011**, *79*, 1061−1078.
(81) Kamisetty, H.; Ovchinnikov, S.; Baker, D. Assessing the Utility of Coevolution-Based Residue-Residue Contact Predictions in a Sequence- and Structure-Rich Era. *Proc. Natl. Acad. Sci. U. S. A.* **2013**, *110*, 15674−15679.
(82) Kabsch, W.; Sander, C. Dictionary of Protein Secondary Structure: Pattern Recognition of Hydrogen-Bonded and Geometrical Features. *Biopolymers* **1983**, *22*, 2577−2637.
(83) Tien, M. Z.; Meyer, A. G.; Sydykova, D. K.; Spielman, S. J.; Wilke, C. O. Maximum Allowed Solvent Accessibilites of Residues in Proteins. *PLoS One* **2013**, *8*, No. e80635.
(84) Towns, J.; Cockerill, T.; Dahan, M.; Foster, I.; Gaither, K.; Grimshaw, A.; Hazlewood, V.; Lathrop, S.; Lifka, D.; Peterson, G. D.; Roskies, R.; Scott, J. R.; Wilkins-Diehr, N. XSEDE: Accelerating Scientific Discovery. *Comput. Sci. Eng.* **2014**, *16*, 62−74.

https://doi.org/10.1021/acs.jcim.1c00449
J. Chem. Inf. Model. 2021, 61, 3058−3073