

Cite this: DOI: 00.0000/xxxxxxxxxx

Inhibitor binding influences the protonation states of histidines in SARS-CoV-2 main protease[†]

Anna Pavlova,^a Diane L. Lynch,^a Isabella Daidone,^b Laura Zanetti-Polzi,^c Micholas Dean Smith,^d Chris Chipot,^e Daniel W. Kneller,^f Andrey Kovalevsky,^f Leighton Coates,^f Andrei A. Golosov,^g Callum J. Dickson,^g Camilo Velez-Vega,^g José S. Duca,^g Josh V. Vermaas,^h Yui Tik Pang,^a Atanu Acharya,^a Jerry M. Parks,ⁱ Jeremy C. Smith,^{d,i} and James C. Gumbart^a

Received Date

Accepted Date

DOI: 00.0000/xxxxxxxxxx

The main protease (M^{pro}) of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) is an attractive target for antiviral therapeutics. Recently, many high-resolution apo and inhibitor-bound structures of M^{pro}, a cysteine protease, have been determined, facilitating structure-based drug design. M^{pro} plays a central role in the viral life cycle by catalyzing the cleavage of SARS-CoV-2 polyproteins. In addition to the catalytic dyad His41-Cys145, M^{pro} contains multiple histidines including His163, His164, and His172. The protonation states of these histidines and the catalytic nucleophile Cys145 have been debated in previous studies of SARS-CoV M^{pro}, but have yet to be investigated for SARS-CoV-2. In this work we have used molecular dynamics simulations to determine the structural stability of SARS-CoV-2 M^{pro} as a function of the protonation assignments for these residues. We simulated both the apo and inhibitor-bound enzyme and found that the conformational stability of the binding site, bound inhibitors, and the hydrogen bond networks of M^{pro} are highly sensitive to these assignments. Additionally, the two inhibitors studied, the peptidomimetic N3 and an α -ketoamide, display distinct His41/His164 protonation-state-dependent stabilities. While the apo and the N3-bound systems favored N_δ (HD) and N_ε (HE) protonation of His41 and His164, respectively, the α -ketoamide was not stably bound in this state. Our results illustrate the importance of using appropriate histidine protonation states to accurately model the structure and dynamics of SARS-CoV-2 M^{pro} in both the apo and inhibitor-bound states, a necessary prerequisite for drug-design efforts.

Introduction

A new coronavirus, named severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), has caused a worldwide outbreak of viral pneumonia starting in December 2019, known as coronavirus disease 2019 (COVID-19), and, currently, treatment options are very limited. Two potentially druggable targets in SARS-CoV-2 and other betacoronaviruses are the chymotrypsin-like protease (3CL M^{pro}) and the papain-like protease (PL^{pro}), which are responsible for cleaving the large polyproteins translated from the viral RNA. M^{pro} has been shown to cleave at least 11 sites of the polyprotein 1ab, recognizing the sequence Leu-Gln↓Ser/Ala, typically denoted P2-P1↓P1', with the cleavage occurring between the P1 and P1' residues, as denoted by the arrow^{1–3}. Because this sequence is not recognized by any known human protease, M^{pro} is an attractive target for SARS-CoV-2-specific antivirals, and previous

^a School of Physics, Georgia Institute of Technology, Atlanta GA 30332, United States.
E-mail: gumbart@physics.gatech.edu

^b Department of Physical and Chemical Sciences, University of L'Aquila, I-67010 L'Aquila, Italy.

^c CNR Institute of Nanoscience, I-41125 Modena, Italy.

^d Department of Biochemistry, Molecular and Cellular Biology, The University of Tennessee, Knoxville. 309 Ken and Blaire Mossman Bldg. 1311 Cumberland Avenue, Knoxville, TN 37996, United States.

^e UMR 7019, Université de Lorraine, Laboratoire International Associé CNRS and University of Illinois at Urbana-Champaign, Vandoeuvre-lès-Nancy, F-54500, France and Department of Physics, University of Illinois at Urbana-Champaign, 1110 West Green Street, Urbana, IL, 61801, United States.

^f Neutron Scattering Division, Oak Ridge National Laboratory, 1 Bethel Valley Rd, Oak Ridge, TN, 37831, United States.

^g Computer-Aided Drug Discovery, Global Discovery Chemistry, Novartis Institutes for BioMedical Research, 181 Massachusetts Avenue, Cambridge, Massachusetts 02139, United States.

^h National Center for Computational Sciences, Oak Ridge National Laboratory, Oak Ridge, TN 37831, United States.

ⁱ UT/ORNL Center for Molecular Biophysics, Biosciences Division, Oak Ridge National Laboratory, TN 37831, United States.

[†] Electronic Supplementary Information (ESI) available: additional analysis. See

DOI: 00.0000/00000000.

attempts at developing them for other coronaviruses focused on it^{1,4-9}.

Crystal structures of M^{Pro} from SARS-CoV, SARS-CoV-2, and other coronaviruses reveal that this enzyme is a homodimer and is structurally conserved among coronaviruses^{7,8,10,11}. M^{Pro} is a cysteine protease that employs a non-canonical Cys-His catalytic dyad with the overall structure of the M^{Pro} dimer illustrated in Figure 1. The individual monomer consists of three domains: domain I (residues 8-101), domain II (residues 102-184), and domain III (residues 201-303); domains I and II interact to create the catalytic site, while the α -helical domain III is responsible for dimerization^{7,8}. Given that earlier work has shown that the M^{Pro} monomer is catalytically inactive¹²⁻¹⁴, it appears that dimerization is essential for M^{Pro} catalytic activity. The first residue, Ser1, of the N-finger (residues 1-7) interacts with Glu166 of the other monomer, shaping and stabilizing part of the substrate binding site¹². The dimerization is abolished by removal of the N-finger and mutation of Tyr161, while the dimerization binding constant is reduced for E166A mutants^{12,14,15}.

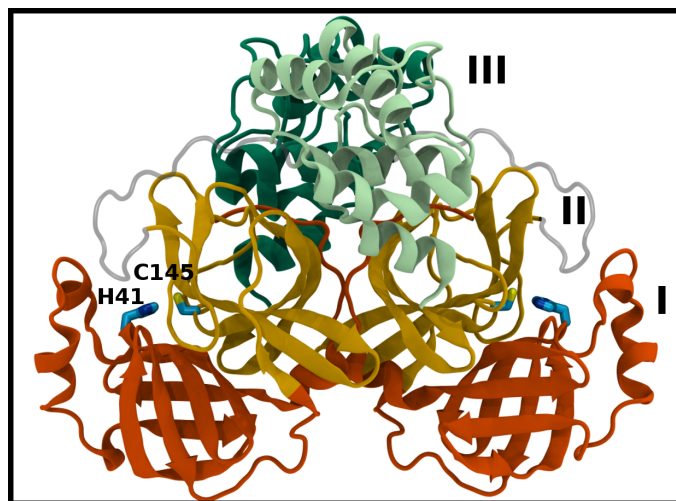


Fig. 1 M^{Pro} dimer structure and binding site interactions (PDB entry 6WQF). M^{Pro} homodimer with the three domains illustrated and color coded as follows: Domain I (dark orange), domain II (gold), and domain III (light green/dark green monomer A/B) with the catalytic dyad residues, His41 and Cys145 (rendered in licorice).

Given the urgent need for effective treatment options for COVID-19, the search for M^{Pro} inhibitors has been intense. For example, several inhibitors have been identified with crystal structures obtained and released, including the peptidomimetic N3 and α -ketoamides (Figure S1)^{7,8}. These inhibitor-bound structures provide excellent starting points for further drug optimization strategies. In addition to the crystal structures of M^{Pro}, both in its apo as well as ligand-bound states^{7,8,11}, recent crystallographic screening¹⁶ has revealed the binding of a multitude of molecular fragments both in the active site as well as at the dimer interface where oligomerization can be disrupted. Such structural studies, complemented by computational tools for drug design, offer tremendous promise for the rapid generation of potent lead compounds for the production of effective

antivirals.

M^{Pro} of coronaviruses employs a His-Cys catalytic dyad for protein cleavage. Mutation of the cysteine to serine reduces catalytic efficiency tenfold, and the enzyme is most active at pH 7.4^{14,17,18}. As observed in neutron crystallographic structures of other proteins, histidines can form hydrogen bonds with main-chain amide nitrogens, resulting in low pK_as^{19,20}. Such histidines will remain in a neutral deprotonated state; for example, in M^{Pro} these include His172 and His246. Several other histidines are positioned near the binding site: His41 (catalytic), His64, His80, His163, His164 and His172, and it has been suggested that protonation and deprotonation of some of these histidines are responsible for changes in enzymatic activity in SARS-CoV^{17,21}. Specifically, crystal structures obtained at pH 6.0 suggest that His163 becomes protonated, causing rearrangement of interactions with the N-finger and partial collapse of the binding site¹⁷. These observations were supported by molecular dynamics (MD) simulations²¹. It was also suggested that in SARS-CoV M^{Pro}, His172 is protonated because both nitrogens of its side chain appear to donate hydrogen bonds (PDB entries 1UK3, 2DUC)¹⁷. For the same protein, MD simulations showed that protonation of His172 increased the flexibility of Glu166 and in turn increased the distance between Glu166 and the N-finger²¹. However, it is not clear how Glu166 flexibility affects enzyme catalysis. Furthermore, released structures of SARS-CoV-2 M^{Pro} show greater agreement between the two monomers than those of SARS-CoV M^{Pro} and none of them show His172 potentially donating two hydrogen bonds^{7,8}. Therefore, to determine the role of the protonation state of His172 in SARS-CoV-2 M^{Pro}, further investigation is needed.

Since the release of SARS-CoV-2 M^{Pro} structures in apo and inhibitor-bound states, multiple MD simulation and docking studies of this enzyme have already been published²²⁻²⁸. Although a precise determination of the specific protonation states of Cys145 and several histidines in M^{Pro} in this pH-sensitive enzyme will be critical to effective and robust computational drug design efforts, these protonation states have not been unequivocally determined. Moreover, recent QM calculations by Poater²⁹ in presence of inhibitors highlight the critical role of His41 acting as a proton shuttle in the catalytic cycle of M^{Pro}. Despite the fact that X-ray crystallography can identify and precisely position non-hydrogen atoms, the low electron density of hydrogen normally precludes the direct X-ray determination of protonation states. Neutron crystallography, on the other hand, can often identify the location of hydrogen atoms even at modest resolution³⁰⁻³², including recently for M^{Pro}³³. However, protonation states in proteins can deviate considerably from intrinsic (i.e., aqueous solution) values due to preferential stabilization of protonated or deprotonated states by interactions with ligands and the protein interior. Furthermore, altering protonation states of titratable groups can modulate protein dynamics and stability via dynamic hydrogen-bonding networks, as has recently been illustrated using a novel graph-based analysis approach for the SARS-CoV-2 S protein^{34,35}.

The protonation states of the catalytic residues, His41 and Cys145, and His164 in SARS-CoV M^{Pro} have also been investi-

gated previously^{18,36}. Mechanistic studies concluded that the catalytic residues are most likely neutral in the active state¹⁸, while MD simulations showed increased distances between the catalytic dyad atoms NE of His41 and S of Cys145 in the charged state in comparison to the neutral state³⁶. However, combined protonation of His164 and deprotonation of Cys145 could not be excluded³⁶. In addition, it was shown that a water molecule actively participates in the cleavage reaction^{36–38}.

Therefore, given similar ambiguity in the protonation states of SARS-CoV-2 M^{Pro}, we have investigated the stability of 12 possible protonation states of this protein. MD simulations of apo and inhibitor-bound M^{Pro} dimers were performed for each state and the resulting structural and dynamical properties were analyzed. Two inhibitors were studied: N3 and one of the most potent α -ketoamides to date (Figure S1), which we refer to simply as ketoamide hereafter. Overall, we conclude that the structural stability of M^{Pro} is indeed affected by the assignment of His protonation states in simulation studies, and the most structurally stable protonation states vary in a ligand-dependent manner.

Methods

Simulations

All simulations used the CHARMM36m force field for proteins and the CHARMM-modified TIP3P water model^{39,40}. Initial CGenFF parameters for the inhibitors N3 and ketoamide were taken from the CGenFF program^{41–43}. Ketoamide charges and bonded parameters that had poor analogies according to the CGenFF scoring system were reoptimized using the force field toolkit (ffTK)⁴⁴; see the supplementary information (SI) for details. Gaussian16 was used for QM calculations required for parameter optimization⁴⁵. N3 parameters were not further optimized because they had lower penalties from CGenFF than ketoamide, and because those penalties mainly involved the methylisoxazole ring, which does not interact with M^{Pro} in the crystal structure.

NAMD 2.13 was used for a set of 20-ns simulations of each potential protonation-state combination^{46,47}. All covalent bonds with hydrogens were kept rigid, allowing integration of the equations of motion with a 2-fs time step. A 12 Å cutoff was used for van der Waals interactions and a smoothing function was applied from 10–12 Å, ensuring a smooth decay to zero. The particle-mesh Ewald method was used to account for long-range electrostatic interactions⁴⁸. Pressure and temperature were kept constant at biologically relevant values of 1 bar and 310 K using a Langevin thermostat and barostat, respectively⁴⁹.

The shorter simulations exploring alternative protonation states were run for 20 ns \times 3 runs for all 12 states for two apo structures, one collected at 100 K (PDB entry 6YB7)⁵⁰ and one collected at room temperature (PDB entry 6WQF)¹¹, as well as for the ketoamide and N3-bound structures (6Y2G⁷ and 7BQY⁸, respectively), giving a cumulative total of 2.88 μ s simulation time. Both 6Y2G and 7BQY are missing the positions of C-terminal amino acids 302–306. Hence, the coordinates for these residues were obtained from two other M^{Pro} structures with intact C-termini. We used the older N3-bound structure (PDB

entry 6LU7) for the C-terminus of 7BQY and the apo structure 6YB7 for the C-terminus of 6Y2G. Although the inhibitors are covalently bound in both crystal structures, we chose to simulate them in their pre-covalent states, prior to bond formation. Previous QM/MM studies have investigated the pre-covalent states using the covalently-bound structures in order to study the cleavage mechanism^{36,38}. Here we used pre-covalent states because we are predominantly interested in the effects of the protonation states on non-covalent binding, which may be masked by covalent attachment. All structures simulated with NAMD were solvated and ionized with 0.15 M NaCl in VMD⁵¹. After minimization, water and ions were equilibrated for 1 ns while the protein and inhibitor, if present, were restrained with a force constant of 2 kcal/mol/Å². In a second equilibration step only the protein backbone was restrained with the same force constant for 4 ns. For analysis, the first 5 ns of the subsequent unrestrained 20-ns simulation runs was discarded.

Extended simulations. The extended simulations focusing on a more limited set of protonation states, in which His41 from 6WQF model was assigned as either HD41 or HE41 paired with HE164, were simulated using Gromacs 2020.1⁵², with five sets of 250-ns simulations run to permit analysis over longer time scales. The molecular systems were built directly from the initial crystal structures, with additional solvent and neutralizing ions added with GROMACS tools. For consistency with the CHARMM36m force field used throughout this study⁴⁰, the same 12 Å cutoff was applied, with particle-mesh Ewald⁴⁸ used for long-range electrostatics. To maintain the temperature at 310 K, a velocity rescaling thermostat was used, with an isotropic Parrinello-Rahman barostat maintaining the pressure at 1 bar⁵³. LINCS was used to restrain the lengths of bonds to hydrogen atoms, enabling 2-fs time steps⁵⁴.

Free-energy perturbation. The relative protein-ligand binding free energies associated to the transformations of residue HD41 into HE41, on the one hand, and residue HE163 into HP163, on the other hand, with either the ketoamide or N3 bound to the enzyme, were determined using free-energy perturbation (FEP)^{55,56}. Towards this end, the protonation state changes were carried out in the apo (unbound) state of the enzyme and in the respective inhibitor-bound state, concomitantly in both monomers of the homodimeric protein. The following protonation states were used for the residues that were not altered in the FEP transformation: HD41, HE164, HE163, HE172 and neutral C145. For both His41 and His172, the transformations were first initiated from the HE state. The change in free energy $\Delta\Delta G^\circ$ was obtained from the difference in alchemical free energies between the bound and unbound states, i.e., $\Delta G_{\text{BAR}}^{\text{bound}} - \Delta G_{\text{BAR}}^{\text{unbound}}$. Considering the nature of the alchemical transformations, the reaction path was stratified into 100 stages of equal widths. Replacement of HD41 and HE163 was performed over 160 ns, both in the bound and unbound states (see Table S4 for individual runs). The dual-topology paradigm was utilized,⁵⁷ whereby the initial and the final states of the alchemical transformation coexist, albeit not interacting. To improve sampling efficiency, geometric restraints were introduced to ensure that the imidazole rings

of the alternate topologies remain superimposed in the course of the amino-acid replacement. At each stage of the stratified reaction path, data collection was prefaced by suitable thermalization in the amount of one-fourth of the total sampling. To augment the accuracy of the free-energy calculation, each alchemical transformation was carried out bidirectionally⁵⁸, and the associated relative binding free energy was determined using the Bennett acceptance ratio (BAR) method⁵⁹. The error bars associated to the relative free energies were computed from the hysteresis between the forward and backward alchemical transformations of the bidirectional FEP calculations, and assume independence of the transformations in the bound and unbound states. All free-energy calculations were performed with the recent single-node path implementation of FEP⁶⁰ available in NAMD 3.0⁴⁷.

Analysis

Root-mean-square deviation (RMSD) of the proteins was measured by comparing the C_α positions to those of the starting structures. This property was calculated for each monomer separately after alignment and aimed at detecting differences in the overall protein structure. The RMSD of the active site was calculated using all non-hydrogen atoms for residues 25 to 28, 38-50, 139-145, 160-176, and 185-195 of the selected monomer and residues 1 and 2 of the other monomer (Figure S3A). We aimed to investigate the changes in the shape of the inhibitor binding site by analyzing this property. Therefore, the RMSD of the site was calculated after alignment of the above-defined active site residues with their positions in the starting structures. The RMSD of the inhibitor was measured by comparing its position over time to its crystallographic position after aligning to the protein. Specifically, this RMSD was calculated after aligning to C_α positions of the dimer complex. All properties were calculated separately for each of the two monomers.

All distribution plots were generated by applying Gaussian kernel density estimation on normalized histograms using the seaborn python package. This estimation was used to generate smoother curves, which are easier to compare and interpret.

Hydrogen bonding analysis was performed with the hydrogen bonding plugin of VMD 1.9.4 with a 3.5-Å and 35-degree heavy-atom distance and angle cutoffs, respectively. Hydrophobic contacts were straightforwardly determined by counting the number of times each ligand carbon atom comes within 4.5 Å of a protein carbon atom. Pocket volume was computed using the Epock VMD plugin with the standard program settings⁶¹. See the SI for definitions and visualization of the binding sites and volumes (Figure S4).

Results and Discussion

The structure of the substrate binding site for the main protease of SARS-CoV-2 includes the catalytic dyad residues (Cys145, His41), as well as recognition pockets for specific substrate residues. These include the S1 pocket, providing His163 for interaction with the highly conserved substrate residue glutamine, and the S2 pocket, which accommodates the hydrophobic P2

leucine. These subsites are often the target for antiviral drug design, and, depending on the length of the inhibitor, additional subsites may be occupied. For example, the long peptidomimetic inhibitor N3 occupies the S1, S2, S4, and S5 sites and extends into the S1' site⁸.

We investigated the effects on the structural properties of the apo and ligand-bound systems by altering the protonation state of the catalytic dyad, Cys145 and His41, as well as those of three histidines near the substrate binding site, His163, His164, and His172. Cys145, His41, and His163 form direct contacts with substrates and inhibitors. His163 HD also forms a hydrogen bond with Tyr161, while His41 is also involved in a network of interactions that includes a water molecule, His164, and Asp187, which in turn is stabilized by a salt bridge with Arg40. In addition, NE of His164 forms a hydrogen bond with the side chain of Thr175. In the structure 6YB7, a hydrogen bond between the His41 side chain and the Gly174 backbone is observed, which is absent in the other three structures studied here. The Arg40-Asp187 interaction bridges domain I with the interface of domains II and III. It has been suggested that the water molecule acts as a third partner in the enzymatic activity of M^{Pro}^{11,36}. Many successful inhibitors^{7,8} of M^{Pro} include a lactam moiety, mimicking the glutamine of the substrate, which binds in the S1 substrate subsite via a hydrogen bond with His163. Maintenance of the S1 pocket⁷ and dimerization state^{14,15,62} have been shown to affect the catalytic efficiency of M^{Pro}. In addition to His163, residues implicated in maintaining either the shape of the S1 pocket or the dimerization state of the protein include Glu166, His172, Phe140 and the N-terminal serine, Ser1', from the other monomer (Figure 2A). For both N3 and ketoamide, a hydrophobic moiety occupies the S2 subsite, which is a relatively deep pocket formed by Met49, Tyr54, Met165, as well as the hydrocarbon portion of the Asp187 side chain^{7,8}. These regions are illustrated in Figure 2B and 2C for the N3 and ketoamide-bound structures, respectively.

Setup of the studied systems

Various methods allow for computation for pKas for specific residues. Most straightforward are webserver, e.g., the Poisson-Boltzmann solver H++⁶³ and the empirical predictor propKa⁶⁴. Importantly, however, while the former predicts HD and HE states of histidine, the latter only distinguishes between protonated and neutral states. Nonetheless, the assignment of histidine protonation states for subsequent detailed studies, e.g., virtual screening, remains challenging⁶⁵. A more accurate, albeit computationally costly approach, is constant pH molecular dynamics (CpHMD), which was recently applied to PL^{Pro} of SARS-CoV-2^{66,67}. Although CpHMD can reveal likely combinations of protonation states, we also wanted to consider the structural consequences of unexpected ones. Therefore, shorter, repeated standard MD simulations with fixed protonation states were used.

The following naming scheme was used for different protonation states: C/CD for neutral/deprotonated Cys145 and HE/HD/HP for histidine protonated on NE, ND, or both ni-

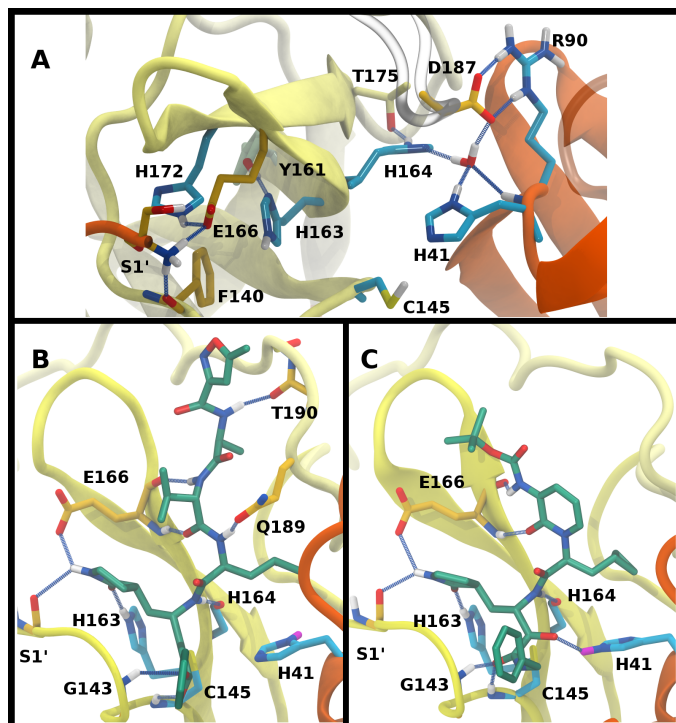


Fig. 2 Hydrogen bonding interactions in the catalytic site for the apo, N3-bound, and ketoamide-bound structures. A) Illustrated for the HD41-HE164 apo state taken from the simulation are the i) catalytic dyad, ii) the crystallographic water bridging His41, His164, and Asp187 as well as His164 with Thr175, and iii) S1 pocket interactions. The N-terminal serine residue is labeled with a prime to indicate that it is from the alternate monomer. B) Binding pocket of N3-bound M^{pro} with hydrogen bonds displayed taken from the HD41-HE164 state. C) Binding pocket of ketoamide-bound M^{pro} with hydrogen bonds displayed, taken from the HE41-HD164 state. The ligands in both B) and C) are rendered in licorice with carbon atoms in green. Note, the His41 hydrogens in the ligand-bound HD41/HE41 states are rendered in magenta, highlighting the alternate protonation for the N3 and ketoamide-bound non-covalent complex with the His164 side chain and water removed for clarity. Hydrogen bonds are illustrated with blue lines.

trogens, respectively. The protonation states of His64, His80 and His246 were not changed from their assigned HE, HD and HE states, respectively, because alternative protonation states were either unlikely or unimportant. Specifically, His64 is solvent exposed, His80 ND forms a hydrogen bond with side-chain oxygen of Asn63, and His246 ND forms a hydrogen bond with backbone NH of Thr243. In addition, the HD state of His163 was not considered because it would prevent hydrogen bonding with the substrates and several known inhibitors^{7,8}. Because simultaneous protonation of ND nitrogens in His41 and His164 was expected to cause sterical repulsion based on available crystal structures we only included one state with simultaneous ND protonation of both residues^{7,8,11}. Table 1 shows the 12 protonation states that were investigated. The protonation states of Cys145, His41 and His164 were considered to be interdependent because of their proximity, while the protonation states of His163 and His172 were assumed to be independent from the other residues. The simulations of these states were performed for two apo crystal structures and two inhibitor bound struc-

tures, one with N3 and one with a ketoamide (see Methods for details).

To facilitate the analysis, the protonation states were divided into three groups, labeled Group 1, 2 and 3, with the goal of determining the most favorable protonation state of Cys145, His41-His164, and His163-His172, respectively. To compare the stability of systems with different protonation states, we computed several properties, namely root-mean-square deviation (RMSD); root-mean-square fluctuations (RMSF); hydrogen bonding; hydrophobic contacts; water occupancy near His41, His164, and Asp187; and volume of the binding pocket. The RMSD was calculated for the C α atoms of the full protein, for the binding site alone, and for the inhibitor relative to the protein. Hydrogen bonding was measured for relevant residues of the protein and between the bound inhibitors and the protein. Residue pairs monitored include the catalytic dyad (Cys145 and His41), those implicated in maintaining the shape of the S1 specificity pocket (His163-Tyr161, His172-Glu166, and the inter-monomer interactions Ser1'-Glu166 and Ser1'-Phe140), and those in regions surrounding the putative catalytic water (Arg40-Asp187). Finally, given the hydrophobic moieties in the N3- and ketoamide-bound M^{pro} structures, particularly the leucine/cyclopropyl group that occupies the S2 subsite, hydrophobic contacts were also measured between the protein and inhibitor. Based on the analysis of these properties, the most structurally stable states were determined.

Apo enzyme

Simulations of the two crystal structures of the apo enzyme state (PDB entries 6WQF and 6YB7, respectively^{11,50}) were performed. Both structural models include full-length M^{pro}, i.e., residues 1-306. The RMSDs, RMSFs, His41(NE) - Cys145(S) distances, and hydrogen bonding patterns are reported in Figures 3 and S5-S7 for the 6WQF simulations and in Figures S8-S11 for 6YB7.

For the 6WQF model, the following states had the lowest RMSDs for both the protein overall and the active site (Figure 3): CD145-HP41-HD164 (Group 1), HD41-HE164 (Group 2), HE41-HE164 (Group 2), and HD41-HP163-HE164 (Group 3). Notably, HE41-HD164-HP172 resulted in increased fluctuations of Ser1, suggesting instability of the N-finger in this state (Figure S5E) and decreased hydrogen bonding capacity (Figure S7C,E). The RMSF profiles (Figure S5) are similar for the different protonation states, and all states showed increased flexibility around residues 49-51 and 192-194, in agreement with previous simulations as well as B-factors for structure 6WQF¹¹.

To better differentiate between low-RMSD states, we evaluated the ability of previously identified residue pairs (Figure 2A) to form hydrogen bonds in each of the 12 protonation states starting from the apo structure (PDB entry 6WQF). Their occupancies are reported in Figure S7. Illustrated in Figures 4 and 5 are persistent hydrogen bonds as well as cases where perturbations of the S1 pocket interactions or hydrogen bonding near the crystallographic water occur. Although many of these interactions have similar propensities, particularly when considering

Table 1 Histidine protonation state combinations considered in this work. Altered protonation states are shown in bold. The corresponding groups and names of the states are also shown

Cys145	His41	His163	His164	His172	Name	Group
C	HE	HE	HD	HE	HE41-HD164	1,2,3
CD	HP	HE	HD	HE	CD145-HP41-HD164	1
CD	HE	HE	HP	HE	CD145-HE41-HP164	1
C	HP	HE	HE	HE	HP41-HE164	2
C	HE	HE	HE	HE	HE41-HE164	2
C	HE	HE	HP	HE	HE41-HP164	2
C	HD	HE	HD	HE	HD41-HD164	2
C	HD	HE	HE	HE	HD41-HE164	2
C	HE	HP	HD	HE	HE41-HP163-HD164	3
C	HE	HE	HD	HD	HE41-HD164-HD172	3
C	HE	HE	HD	HP	HE41-HD164-HP172	3
C	HD	HP	HE	HE	HD41-HP163-HE164	3

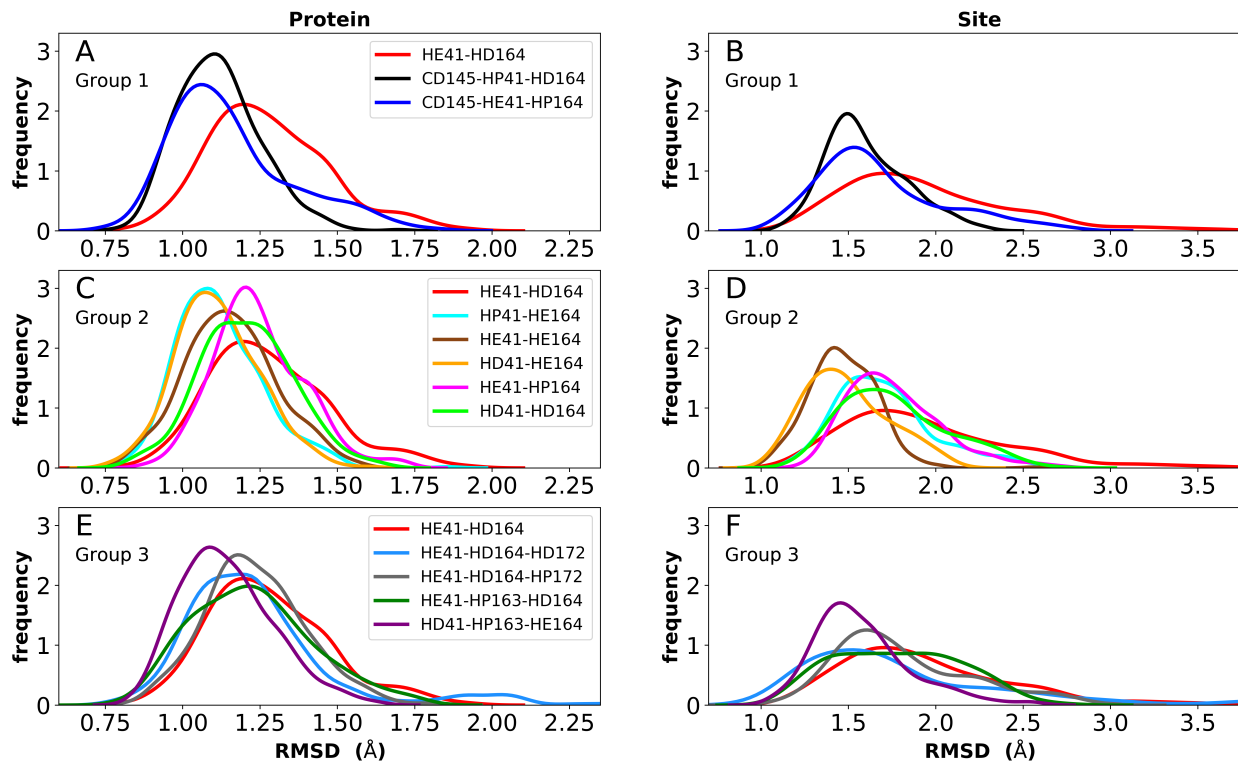


Fig. 3 RMSD distributions for the three protonation-state groups from simulations of the apo structure (PDB entry 6WQF¹¹). A-B) RMSD of (A) protein and (B) active site for Group 1. C-D) RMSD of (C) protein and (D) active site for Group 2. E-F) RMSD of (E) protein and (F) active site for Group 3.

the standard deviations, we used hydrogen bonding involving several specific interactions in the S1 pocket and the catalytic dyad (Figure S7) to assist in eliminating protonation states.

Interactions with the S1 specificity pocket. For states with a neutral His163, the His163-Tyr161 side chains are engaged in a hydrogen bond with the histidine/tyrosine acting as the acceptor/donor, respectively (Figure 2A). This result is in agreement with known hydrogen bonding propensities for tyrosine residues⁶⁸. The most dramatic differences occur upon protonation of His163 or His172 (Figures 4 and S7). For the single HP172 state considered, there is a significant reduction in the hydrogen-bonding occupancy for the His163-Tyr161 pair (Figures 4B and S7A). Moreover significant separation of the Ser1' and Glu166 residues is observed, resulting in a disruption of

this S1 pocket interaction (Figures 4B and S7C). This is consistent with the structural instability of the N-finger region in the HE41-HD164-HP172 state discussed above. When His163 is protonated (the HP163 states), the S1 pocket His163-Tyr161 interaction is ruptured due to loss of the hydrogen bonding acceptor moiety in the charged histidine side chain (Figures 4C and S7A). Only in the HD41-HP163-HE164 system does the tyrosine(acceptor)/histidine(donor) pair occur; however, the occupancy for this interaction (<20%) is quite low. Furthermore, significant perturbation of the hydrogen bonding between His172 and Glu166 is also observed in the charged HP163 states (Figures 4C and S7B). This finding is in accord with the suggestion by Tan et al.²¹ that protonation of His163 will contribute to altering the properties of the S1 specificity pocket. Given the dis-

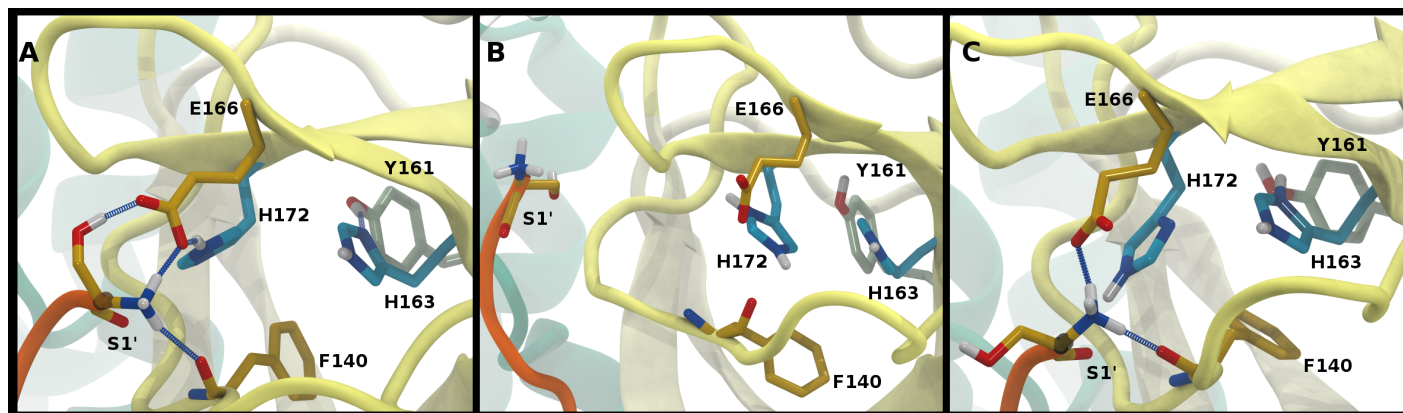


Fig. 4 Hydrogen bonding in the S1 pocket. Example configurations of A) HD41-HE164 characteristic of robust S1 pocket interactions, B) HE41-HD164-HP172 illustrating the rupture of the S1'-Glu166 interaction and loss of the His163-Tyr161 hydrogen bond, and C) HE41-HP163-HD164 depicting the loss of the Tyr161 hydrogen bond donation and the His172-Glu166 interaction.

ruption of the hydrogen bonding just discussed for HP163, the low-RMSD state from Group 3, HD41-HP163-HE164 may be discarded. Also of note is the observation that the S1 pocket interactions are significantly disturbed in the HD172 state (HE41-HD164-HD172; Figure S7) reinforcing the removal of this set of protonation states.

Catalytic Dyad (Cys145, His41). The NE-S distance between the two catalytic residues is centered around 4 Å for most states with neutral Cys145 and is slightly reduced for the states with deprotonated Cys145 (CD145) and/or charged HP41 (Figure S6B,D,F). Moreover, direct hydrogen bonding between the catalytic dyad residues Cys145 and His41 is rare, with non-zero occupancies only for the protonated systems, CD145-HE41-HD164 and CD145-HP41-HD164, at 22% and 26%, respectively (Figure S7F). These results are in general accord with the observed Cys145-His41 distance distributions discussed above and the elongated/weaker hydrogen bonding propensities of sulfur atoms⁶⁹. Given the longer distances seen in the crystal structures and the shifted NE-S distances in the CD145/HP41 simulations we have removed the zwitterionic CD145-HP41-HD164 (Group 1) state from further consideration.

Crystallographic water. Interactions with water are particularly relevant for His41, which interacts with both His164 and Asp187 via a bridging crystallographic water (Figure 2A). This water molecule has been suggested to play a role in the catalytic reaction by stabilizing the positive charge accumulated on His41 and by assisting in proton transfer from Cys145 to His41^{11,36}. Given the possibility of water exchange, we have computed the occupancy of the putative catalytic water site by obtaining the fraction of frames in which a water molecule is simultaneously within 3.5 Å of His41(ND1/NE2), His164(ND1/NE2), and Asp187(OD1/OD2). The average occupancy is substantially above 80% for only the HD41-HE164 and HD41-HD164 states (Table S3), suggesting a particularly stable structural arrangement of His41, His164, Asp187, and water. It is noteworthy that in the HD41-HD164 state, the His164 changes conformation; the resulting geometry is analogous to that observed for the HD41-HE164 state. This similarity is clearly illustrated

in Figures 5A and B, where the NE nitrogen on His164 has been colored magenta to highlight this rotation. This alteration of conformations produces the increased water occupancy observed in the HD41-HD164 state (Table S3) and is consistent with the increased site RMSD distribution observed for this system relative to HD41-HE164 (Figure 3D). Lastly, illustrated in Figure 5C is an example from the HE41-HE164 state, in which the crystal structure arrangement of the water molecule is lost, and the water interacting with His41 adopts an alternate orientation, often still interacting with HE164.

Rotation of the side chain of His164 is characteristic of the HD states and generates additional state dependent hydrogen bonding propensities for this residue. We observe that in the HE and HP protonation states of His164 (Figure 5A), a hydrogen bond is present between the side chains of His164 (NE protonated and donating) and Thr175 (accepting). In contrast, for the His164 HD states the proton on the ND nitrogen is involved in a hydrogen bond with the backbone carbonyl of Met162 as illustrated in (Figure 5B). This occurs as a result of a rotation of the His164 side chain. Finally for 11 of the 12 states studied, the side chain of Thr175 donates a hydrogen bond to the main chain carbonyl of Asp176 with occupancy greater than 75% (Figure 5). The only exception is the HE41-HP163-HD164 state where the Thr175 side chain donates a hydrogen bond to HD164 (~50%).

Pocket volume. We also examined the pocket volume changes (PDB 6WQF: Figure S6A,C,E and PDB 6YB7: Figure S10A,C,E). For most states, the pocket volume varies between 200 and 600 Å³. Slight increases in pocket volume are observed for states HE41-HD164 and HE41-HP164 in comparison to other systems. However, it is not clear what pocket volume is optimal for substrate binding. A comparison between the recently published room temperature crystal structure of M^{pro}¹¹ and the N3-bound crystal structure indicate regions that undergo significant conformational changes upon ligand binding. Of note, these regions include residues near the P2 site (residues 49-50), the P3-P4 site (residues 166-170), and the P5 loop (residues 190-194) and imply an induced-fit type of ligand-protein interac-

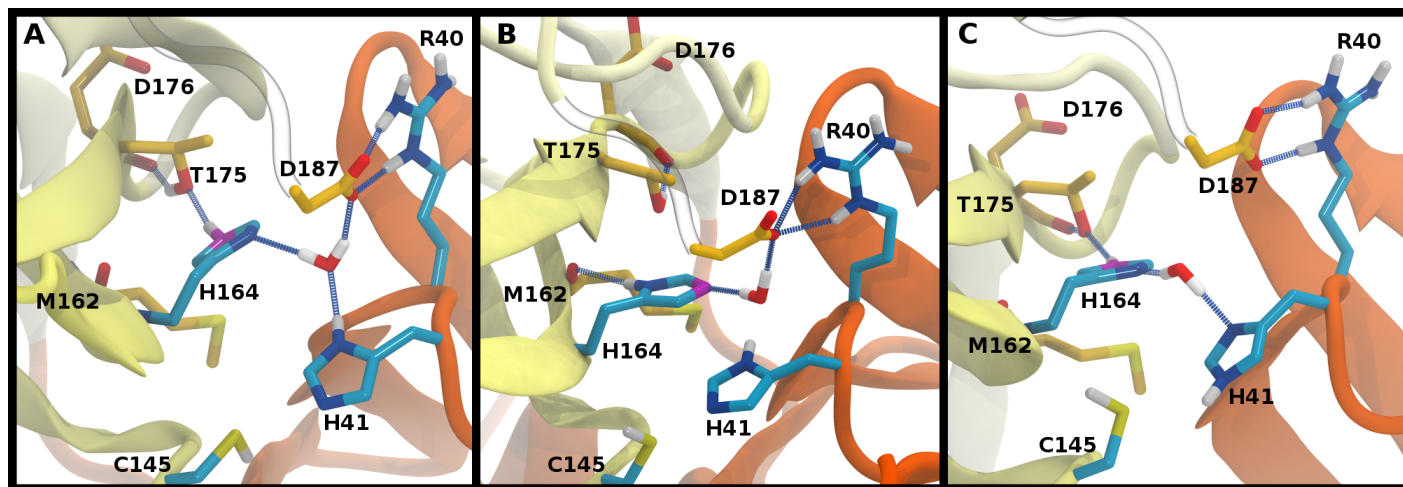


Fig. 5 Hydrogen bonding involving the crystallographic water and His164 local environment. In each image, the NE nitrogen in the His164 side chain is colored magenta. A) HD41-HE164 illustrating hydrogen bond donation from His164 to Thr175. B) HD41-HD164 illustrating the His164 side chain rotation such that a hydrogen bond is formed with the backbone carbonyl of Met162. C) HE41-HE164 illustrating the disruption of the hydrogen bonding interactions and loss of the interaction of the water with Asp187.

tion. In accord with these previous results, our RMSD/RMSF-per-residue results (Figure S5) reveal a high degree of flexibility/plasticity in these same regions, suggesting that the ligand site volume fluctuations we observed are a signature of active site flexibility.

Simulation of the apo state starting from PDB 6YB7. In addition to the room temperature M^{pro} structure¹¹, a low temperature (100 K) crystal structure has been released⁵⁰. The distinguishing feature in the vicinity of the active site is the altered rotational state of His41. Given the critical functional role His41 plays in the enzymatic activity of M^{pro} , we have explored possible differences in the structural stability of M^{pro} upon protonation state variations (Table 1) using 6YB7 as the starting structure. In general, the behaviors observed in the simulations based on PDB entries 6WQF and 6YB7 are strikingly similar, including the behavior of the His164 side chain. The protonation state with low RMSD for both the protein and the catalytic site residues is HD41-HE164 (Figure S8). However, there are several differences; for example, the states CD145-HP41-HD164 and HE41-HE164 appeared to be less structurally stable based on the RMSD in the simulations starting from the 6YB7 structure and in contrast to the simulations based on 6WQF, the His164 in HE41-HP163-HD164 rotates and forms a hydrogen bond with the backbone of Met162. The RMSFs (Figures S5 and S9 for the 6WQF and 6YB7 simulations, respectively), hydrogen bonding (Figures S7 and S11), and water occupancy (Table S3) are mostly the same. There is a reduction in the water occupancy for the HD41-HE164 system in the 6YB7 system; however, the standard deviation is quite large (Table S3).

Collectively, based on RMSD, RMSF, and hydrogen-bonding occupancies in the aforementioned simulations, the most likely apo M^{pro} states appear to be HD41-HE164 and HE41-HE164. However, it should be noted that the differences between states in Group 2 are small, and, therefore, other protonation states of His41-His164 are also possible.

Extended simulations. To provide further differentiation between the neutral HD and HE states for His41 and His164, we performed five 250-ns simulations of states HD41-HE164 and HE41-HE164, starting from the room-temperature 6WQF structure. The RMSD of each protein shows that the overall structure is highly structurally stable over the course of the simulations for both states (Figure 6). The RMSD increases slightly from the shorter simulations; however, it remains below 3 Å. In contrast, significantly larger RMSDs were observed when only the active site was considered, particularly for the state HE41-HE164. Examination of five separate runs shows frequent variations in site RMSD over the course of 250 ns for almost all runs (Figure S12), suggesting that the active site is more flexible than the overall protein structure, as has been seen in previous MD simulations^{27,28}. The RMSF and RMSD per residue (Figure S13) have similar shapes to those from the shorter apo simulations (Figures S5 and S9). Large variations in the pocket volumes were also observed, ranging from 0-800 Å³, indicating high site flexibility. Figures S3B,C compare the active sites with small and large volumes and show that the volume is affected by the position of Met165 inside the pocket and by movements of two small loops consisting of residues 47-50 and 188-190 at the pocket opening. These loops correspond to the two largest peaks in the RMSF analysis (Figure S13).

We monitored hydrogen bonding to residues in the S1 pocket (Figure S14). Overall, the two simulations give similar results, with the HE41-HE164 system having a slightly enhanced occupancy for four of the five interactions. Similar to the shorter apo state trajectories, there is a hydrogen bond between the His164 (donating) and Thr175 (accepting) side chains in the HD41-HE164 state; it is also present in the HE41-HE164 states, but to a much lesser extent, with an occupancy of ~20%. In both systems, the Thr175 hydroxyl side chain interacts with the Asp176 backbone carbonyl, accepting a hydrogen bond. No hydrogen bond between Cys145 and His41 was present in either

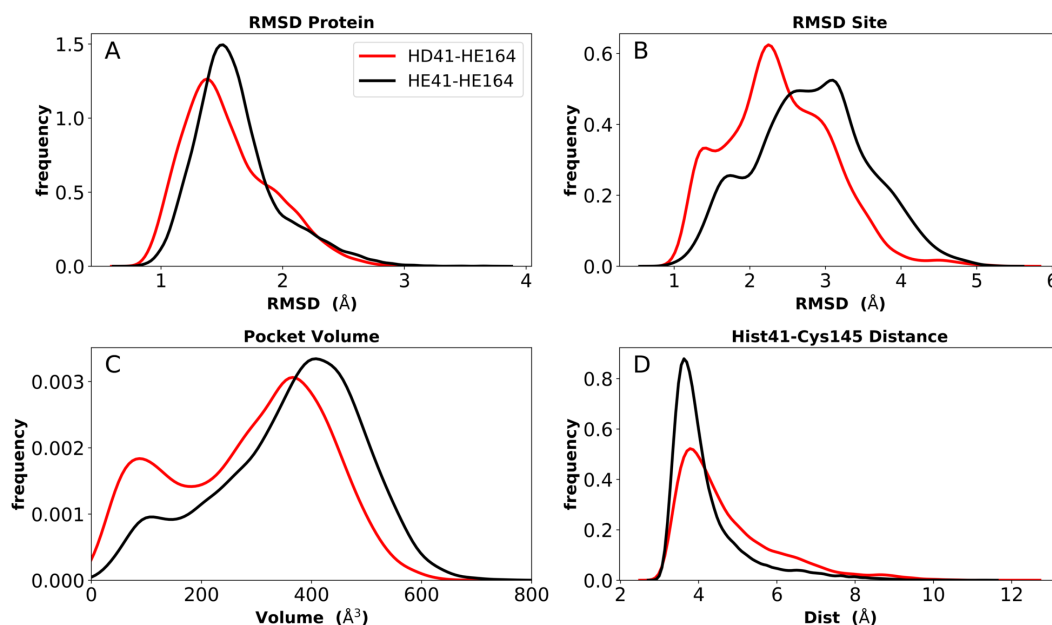


Fig. 6 Analysis of longer MD simulations for states HD41-HE164 and HE41-HE164. A) RMSD of protein only. B) RMSD of active site. C) Pocket volume. D) Distance between NE of His41 and S of Cys145.

system, and the distance between NE of His41 and S of Cys145 remained at ~ 4 Å (Figure 6D).

Also analyzed for these longer runs was the interaction of the His41-His164-Asp187 triad with the catalytic water. For HE41-HE164 in all five trajectories for both monomers, and similar to the shorter apo runs, this arrangement was lost with water forming this trio of interactions for only 21% ($\pm 17\%$) on average, while for HD41-HE164 this percentage increased to 53% ($\pm 28\%$). Moreover, the average percentage for a specific water molecule residing in this region, increases from 4% ($\pm 4\%$) for HE41-HE164 to 36% ($\pm 30\%$) for HD41-HE164. Although the standard deviations are large, these results suggest that, relative to HE41-HE164, the HD41-HE164 state has a greater propensity to accommodate a water molecule at this site. Therefore, after taking the longer simulations into account, we propose that the HD-HE164 state is the most likely apo M^{pro} protonation state.

N3-bound state

The structure of N3, previously developed to inhibit the main protease of multiple coronaviruses, bound to SARS-CoV-2 M^{pro} has been determined (PDB entry 7BQY)⁸. Given that the S1 specificity site for M^{pro} has a nearly absolute requirement for glutamine at position P1 for protein cleavage to occur⁷⁰, it is not surprising that the lactam moiety fits well into this site, forming hydrogen bonds with both His163 and Glu166 (Figure 2B). The overall orientation of the ligand is guided by backbone hydrogen bonds between Glu166 and the peptide, as well as the Gln189 side chain. Hydrophobic contacts are also present, as the leucine side chain is inserted into the S2 subsite, a region known to accommodate non-polar substrate groups⁶. Overall, these diverse interactions maintain N3 in an orientation competent for the inhibition of M^{pro} enzymatic activity. In light of these sta-

bilizing interactions, we monitored the RMSD of the protein, site, and inhibitor (Figure 7), as well as the RMSF/RMSD per residue, pocket volumes, and hydrogen bonding/hydrophobic contact propensities (Figures S15-S18) across the 12 systems.

In contrast to the apo systems, RMSD measurements of the inhibitor-bound systems are more sensitive to changes in the protonation states. All states in Group 1 induce high RMSD of either protein and site or the inhibitor, while out of the Group 3 states, only HE41-HP163-HD164 has a low RMSD for these quantities (Figure 7). In addition, for Group 1 states, the distance between Cys145 S and His41 NE was significantly decreased for the state CD145-HP41-HD164 and significantly increased for the state HE41-HD164. (Figure S16B,D,F). This trend was not observed for apo systems. Slightly increased pocket volumes were observed across all systems relative to the apo states (Figure S16A,C,E). In the N3-bound simulations, the total hydrogen bonding interactions and hydrophobic contacts between the ligand and protein were very similar across all the trajectories (Figure S17). The behavior of the hydrogen bonding of the S1 subsite residues is similar to that in the apo simulations in which protonation of His163 ruptures the interaction with Tyr161 (Figure S18). In line with the apo results, the His172-Glu166/His164-Thr175 hydrogen bond is lost in the HD172/HD164 states, respectively, with the Thr175 side chain donating a hydrogen bond to the backbone of Asp176.

Multiple protonation states from Group 2 appear to be feasible (Figure 7) although increased RMSDs were observed for HP41-HE164 and HE41-HE164, and a small decrease in N3-protein hydrogen bonding was observed for HD41-HD164 (Figure S17). Therefore, we conclude that the remaining states in Group 2, namely HD41-HE164, HE41-HP164 and HD41-HD164, are the most structurally stable N3-bound states. Given that

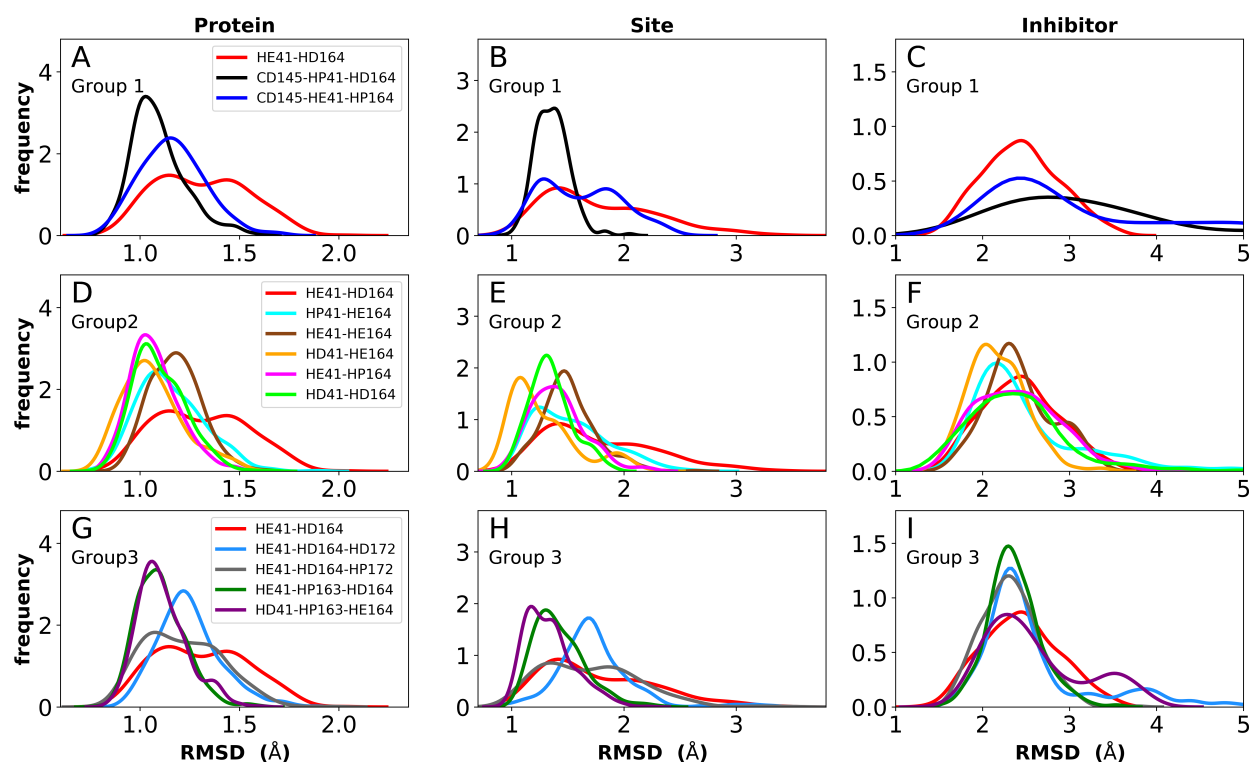


Fig. 7 RMSD distributions for the three protonation state groups from simulations of the N3-bound structure (PDB entry 7BQY⁸). A-C) RMSD of (A) protein, (B) active site, and (C) inhibitor for Group 1. D-F) RMSD of (D) protein, (E) active site, and (F) inhibitor for Group 2. G-I) RMSD of (G) protein, (H) active site, and (I) inhibitor for Group 3.

HD41-HE164 has a lower RMSD (Figure 7E) for the site residues and the inhibitor (Figure 7F) than either HD41-HD164 or HE41-HP164, and also has robust hydrogen bonding and hydrophobic contacts, we propose that this state is the most favorable of the three. These interactions, central to the formation of the non-covalently bonded complex, are explicitly illustrated in Figure 2B.

Ketoamide-bound state

The ketoamide-bound M^{pro} structure (PDB entry 6Y2G) was also simulated in the 12 protonation states listed in Table 1. In contrast to the peptidomimetic N3, the ketoamide carbonyl can accept a hydrogen bond from His41. As illustrated in Figure 2C, this bond occurs only when the NE nitrogen carries a proton. Thus, the ketoamide-bound systems display increased sensitivity to the protonation state of His41 compared to the N3-bound systems. The RMSDs of the protein, site, and inhibitor are illustrated in Figure 8, with RMSF, hydrogen bonding, and hydrophobic contact propensities across the 12 systems also reported (Figures S19-S22).

While the RMSF (Figure S19) profiles and hydrogen bonding of the S1 specificity pocket residues (Figure S22) are similar to the apo and N3-bound simulations, notable differences appear in the RMSD and hydrogen bonding distributions. For Group 1, both states with a deprotonated Cys145 exhibit relatively large and shifted RMSD distributions for the inhibitor (Figure 8C), while the HD41 and HP41 states in Group 2 all dis-

play a wide distribution with a slight shoulder at larger values (Figure 8F). These shifts indicate some level of instability of the ligand in the binding pocket. Lastly, of the Group 3 states, in addition to HE41-HD164 only HD41-HP163-HE164 displays an unshifted site and inhibitor RMSD (Figures 8H,I). Of note, for the HD41-HP163-HE164 system, there is a modest reduction in the hydrogen bond count between the ligand and protein (Figure S21). Thus, we conclude that the ketoamide ligand requires a neutral Cys145 and protonation on the NE nitrogen of His41, leaving HE41-HE164, HE41-HD164, and HE41-HP164 as possible protonation states for the ketoamide-bound M^{pro}. In line with the apo and N3 simulations, a hydrogen bond between the His164 (donating) and Thr175 side chains is present only in the HE164 and HP164 states, while the His164 side chain frequently rotates in the HD164 states to form a hydrogen bond with the backbone of Met162. This rotation is analogous to that observed in the HD164 apo simulations (Figure 5). Here, we observe that the hydrogen bond to the ketoamide is possible in either conformation of the HE41-HD164 state (Figure 9). In contrast to the other HD164 states, His164 in HE41-HD164 and in HE41-HP163-HD164 also accepts a hydrogen bond from Thr175 (~20% of the time).

Notably, protonation states that have low RMSD in N3-bound and apo structures, e.g., HD41-HE164 and HE41-HP163-HD164, showed increased displacement of the ketoamide inhibitor, as measured by its RMSD (Figure 8) and significantly reduced total hydrogen bonding to M^{pro} (Figure S21). In partic-

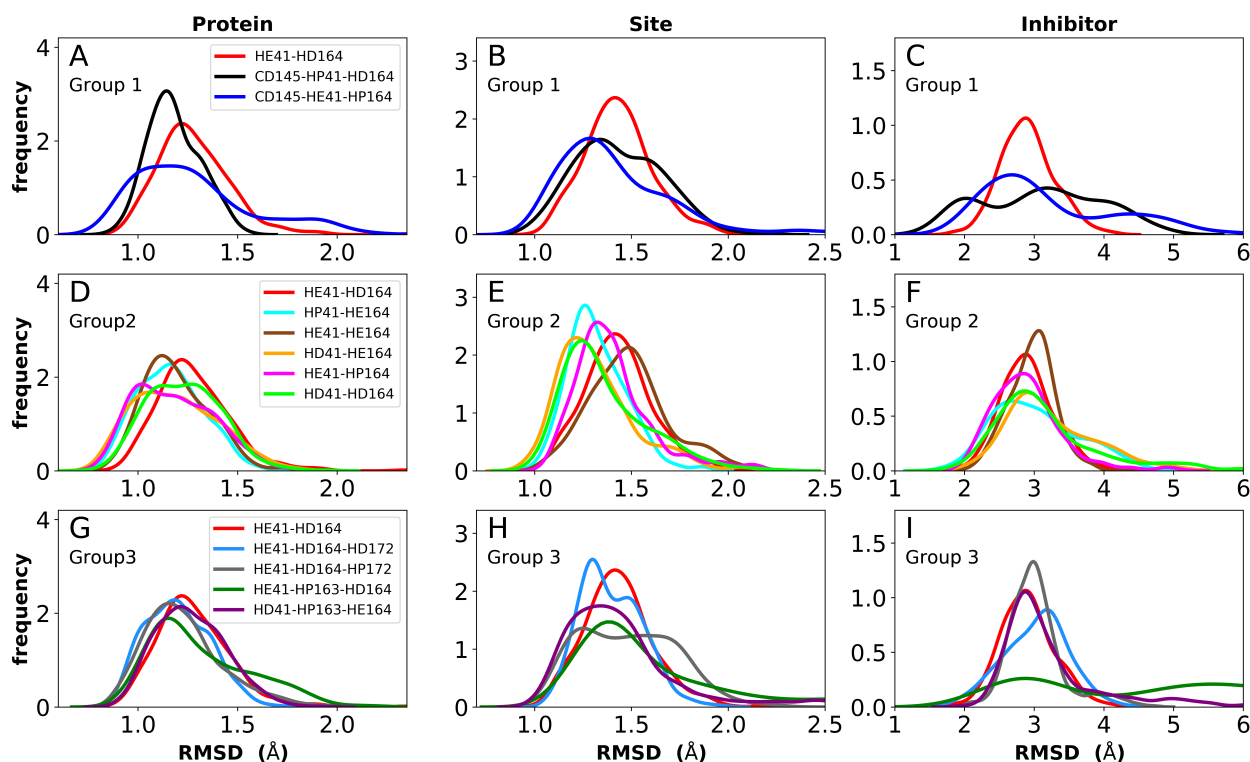


Fig. 8 RMSD distributions for the three protonation state groups from simulations of the ketoamide-bound structure (PDB entry 6Y2G). A-C) RMSD of (A) protein, (B) active site, and (C) inhibitor for Group 1. D-F) RMSD of (D) protein, (E) active site, and (F) inhibitor for Group 2. G-I) RMSD of (G) protein, (H) active site, and (I) inhibitor for Group 3.

ular, the most dramatic effect observed is for the HE41-HP163-HD164 system. Although the HD41-HE164 state still has low RMSD for both the protein and the active site, increased RMSD of inhibitor and decreased inhibitor-protein hydrogen bonding indicates that this protonation state is structurally unstable for ketoamide-bound M^{pro}. This observation is in contrast to apo and N3-bound simulations, where the HE41-HD164 state was not particularly stable structurally instead, and appeared to be destabilizing for the N3-bound state due to high RMSD. This is most likely a result of the sensitivity of the ketoamide-bound simulations to the presence of the HE41 carbonyl hydrogen bond, not present in the N3-bound structure (Figure 2B,C).

Free-energy calculations for His41 and His163 protonation states

As observed in the preceding discussion of possible protonation states, the state of His41 is quite sensitive to the presence of the ligand, favoring HD41 in the N3-bound system and HE41 in the ketoamide system. In order to further delineate the relative stability between HD41 and HE41 in the two systems, free-energy perturbation calculations (FEP) were performed (Table 2). In addition, we investigated the change in free energy for HE163→HP163. The FEP calculations were only used for the two histidines that directly interact with the bound ligands. Our results confirmed that while HE41 is preferred by ketoamide by 1.3 kcal/mol, the HD41 state is preferred by N3 by 0.8 kcal/mol (both numbers per monomer). Although small (1-2 kT), the en-

ergetic differences for the two states are statistically significant (Table 2). In addition, the difference in structural stability for the two states is also supported by changes in inhibitor RMSD and hydrogen bonding with the protein.

Although the HE163→HP163 transformation was unfavorable for both inhibitors (Table 2), the magnitude of free energy change was larger for ketoamide by 1.5 kcal/mol. This is in qualitative agreement with our analysis of hydrogen bonding for the two compounds in HP163 states, which shows a larger loss in hydrogen bonding for ketoamide (~2.5 hydrogen bonds lost on average) than in the corresponding N3 state (~0.5 hydrogen bonds lost; Figures S17 and S21) relative to the HE41-HD164 state.

Table 2 The relative protein-ligand binding free energy, $\Delta\Delta G^\circ$, corresponds to the difference in binding affinities between the initial and the target (WT) enzyme, i.e., $\Delta G_{\text{tar}}^\circ - \Delta G_{\text{init}}^\circ$, which is equal to the difference in alchemical free energies between the bound and unbound states, i.e., $\Delta G_{\text{BAR}}^{\text{bound}} - \Delta G_{\text{BAR}}^{\text{unbound}}$ (see Table S4). The $\Delta\Delta G^\circ$ values reflect the concomitant amino-acid protonation state change in both monomers of the homodimeric enzyme. The values supplied in parentheses represent the relative protein-ligand binding free energy per monomer. All energies are in kcal/mol.

transformation	substrate	$\Delta\Delta G^\circ$
HD41→HE41	ketoamide	-2.7±0.3 (-1.3±0.3)
HE163→HP163	ketoamide	+6.7±1.4 (+3.4±1.4)
HD41→HE41	N3	+1.5±0.1 (+0.8±0.1)
HE163→HP163	N3	+3.9±1.0 (+1.9±1.0)

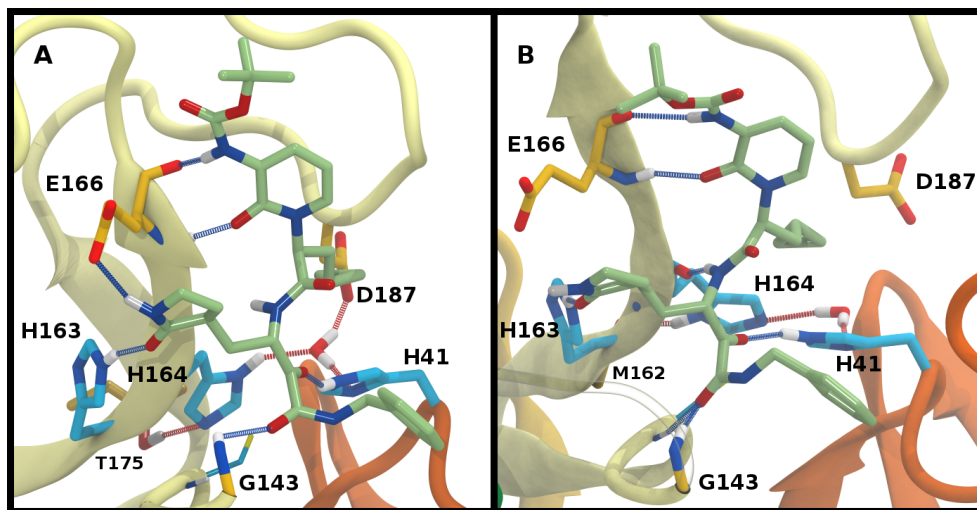


Fig. 9 Ketoamide hydrogen bonding in the HE41-HD164 protonation state. In both panels, hydrogen bonds between the ligand (light green licorice) and the protein are indicated with a blue line, while those with water or between protein residues are red. A) Region around the crystallographic water. B) conformation in which the His164 has rotated, making a hydrogen bond with the backbone of Met162. The loss of the interaction with Asp187 is illustrated.

Conclusions

To combat COVID-19, there is tremendous effort aimed at developing both vaccines and antiviral drugs against its causative agent, SARS-CoV-2. The main protease of SARS-CoV-2 (M^{pro}), a homodimeric cysteine protease, has emerged as an attractive target for drug design given 1) its critical role in early stages of the virus replication cycle, 2) its similarity to main proteases from other betacoronaviruses, thereby leveraging earlier antiviral development efforts, and 3) its unique substrate specificity, cleaving primarily after a glutamine, a target unknown for other host cell proteases^{5–7}.

A number of crystal structures of M^{pro} have been determined, including the apo state as well as bound to covalent, non-covalent, and fragment-based inhibitors^{7,8,11}. The substrate binding site in M^{pro} consists of a catalytic Cys145-His41 dyad, as well as several histidine residues in close proximity to the catalytic site including His172, His163, and His164. These residues, along with Glu166 and the N-terminus of the other monomer, Ser1, form an interlocking hydrogen-bonded pocket, which is the target of most designed inhibitors. For computational drug design strategies to be effective, the details of the structure and dynamics of these residues are required, particularly under different protonation states. In fact the sensitivity of computational results to histidine protonation state choices is a common occurrence in computational drug design. For example, G protein coupled receptors (GPCRs) are a frequent target comprising a large market share⁷¹. A recent computational study of relative binding free energies for two GPCRs, adenosine 2A and orexin-2, has illustrated that alternative histidine tautomeric/protonation states impacts the overall stability of the bound ligand and thus contributes to the overall performance of relative free energy predictions⁷². Here we have enumerated these states and determined the effects of protonating various residues for M^{pro} in both the apo form as well as ligand-bound

complexes using MD simulations.

In the present study, we have demonstrated that the combination of protonation states for histidines in or near the catalytic site can have a profound impact on M^{pro} 's structural stability, as measured by RMSD and RMSF, as well as the hydrogen bonding occupancies, hydrophobic contacts, and catalytic subsite characteristics, e.g., available pocket volume. Examining these properties, we conclude that the protonated His41/deprotonated Cys145 state of the catalytic dyad is structurally unstable in the crystal structure conformations as exemplified by increased RMSD, altered hydrogen bonding patterns, and unbinding of the inhibitors. However, this state may exist as a transient reaction intermediate, as has been proposed by previous computational studies^{29,36}.

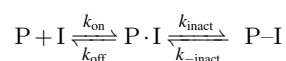
The most likely histidine protonation state for each structure was also determined (Table S5). It was shown that His163 and His172 protonation states other than HE result in significant perturbations to several hydrogen bonds compared to crystal structure conformations. Although HE protonation was expected for these two residues based on structural data, concurrent HD protonation could not be excluded without the results from our MD simulations, particularly for His172, which is thought to be protonated in M^{pro} of SARS-CoV¹⁷. In addition, decreased pocket volume was frequently observed in the simulations for the HP163 state; free-energy calculations showed decreased affinity for both inhibitors in this state, which qualitatively tracks the hydrogen bonding reductions observed. The sensitivity of the M^{pro} structure to the protonation states of His163 and His172 may explain why M^{pro} is inactive at low pH^{18,21}. At low pH, HP163 is expected to interfere with substrate binding due to decreased pocket volume and substrate affinity, whereas fluctuation of the N-finger and decreased hydrogen bonding within the protein, induced by the HP172 state, could interfere with catalysis.

In contrast to the other histidines, we determined that multi-

ple protonation combinations are feasible for the His41-His164 pair. It is possible that these residues adopt different protonation states during protein cleavage. We conclude that HD41-HE164 appears to be the more stable state for the apo and N3-bound structures, whereas HE41-HD164 is more structurally stable for the ketoamide-bound structure. This change in protonation state preference of M^{Pro} for the two inhibitors was also confirmed with free-energy calculations (Table 2). Unlike N3, ketoamide has two carbonyl groups around its reactive site, facilitating additional bonding to M^{Pro} (Figure 2C). In fact, the structural data indicate that protonation of the NE nitrogen in His41 is preferred in order to provide an interaction with the ketoamide carbonyl. Consequently, ketoamide has slightly better in vitro inhibition activity than N3 (5 μ M vs. 17 μ M)^{7,8}, although different reaction mechanisms of covalent bond formation with Cys145 for Michael acceptors and ketoamides as well as their ability to go through membranes can also contribute to differences in activity. Nevertheless, additional optimization to improve their potency is desirable for both compounds.

Previous simulations of M^{Pro} used either Pymol, H++ web-server, or Protein Preparation Wizard from Schrödinger in order to assign hydrogens to the histidines and to Cys145^{22–28}. To compare automatic protonation assignment with our results, we have examined the protonation states obtained from the H++ server and from Protein Preparation Wizard at pH 7.0 (Table S5)^{63,73}. Because the H++ server did not include the ligands in the output, we only present H++ results for the apo structures⁶³. Both methods predicted that the catalytic dyad was neutral, and His172 was in the HE state in all structures, in agreement with our results. However, differences from our MD results were observed for histidines 41, 163 and 164 (Table S5). In particular, the change in protonation states in the presence of ketoamide was not predicted by Protein Preparation Wizard. Although automatic protonation assignment programs greatly assist in protein structure preparation, visual inspection is also recommended. Histidines, having a pKa near physiological pH as well as nearly isoenergetic tautomers, are expected to have protonation state patterns that display high sensitivity to the local environment. As demonstrated here, this is very true for M^{Pro}, in which the histidines are part of an extended hydrogen-bonded network and undergo ligand-dependent protonation state modulation. In other situations such as this, we expect that the stability analysis developed here will be particularly valuable for determining an optimal protonation pattern.

Finally, we include a note on the particular role of the catalytic residue His41. Covalently bound ligands elicit their inhibition by a two-step mechanism⁷⁴, starting when a prereactive non-covalently bound complex forms (P·I) between the protein (P) and the inhibitor (I) governed by an equilibrium constant $K_i = k_{\text{on}}/k_{\text{off}}$. Subsequently, the reaction between the target protein residue, in this case the nucleophilic Cys145, and the ligand occurs, producing the covalently bonded adduct (P–I). In the case that k_{inact} is much larger than $k_{-\text{inact}}$, the covalent binding is essentially irreversible.



For M^{Pro}, the critical role of His41 and its protonation is twofold. First, it provides important hydrogen-bonding interactions in the formation of the prereactive, non-covalently bonded complex. Second, His41, as part of the catalytic dyad, is recognized as an active participant in covalent bond formation, whose contribution to the catalytic reaction has recently been the focus of QM/MM calculations^{29,38}.

Here we have demonstrated the importance of protonation state propensities of His41 in the formation of the non-covalently bonded prereactive complex. Its role in covalent bond formation has been investigated in previous Density Functional Theory calculations, which have shown that the His41 HE and HD protonation states have similar energies, and that proton transfer from Cys145 to His41 is feasible for both histidine states, supporting our conclusion here²⁹. In both protonation states, His41 is expected to be crucial for the covalent reaction by accepting a proton from Cys145^{29,36}. Side chain rotation of His41 is frequently observed for both HE and HD states, and in presence of both inhibitors, in agreement with previous a QM/MM study by Poater²⁹, allowing for either ND or NE nitrogen to subtract a proton from Cys145. Therefore the different protonation states of M^{Pro} histidines will need to be considered for optimization efforts of N3 and ketoamide, as well as for the rational design of other inhibitors. We expect that other peptidomimetic Michael acceptors and α -ketoamides will have the same protonation state preferences as N3 and ketoamide, respectively. However, different warheads may produce alternate stable protonation states, further complicating drug design and requiring a stability analysis as described here for N3 and ketoamide. In silico high-throughput screens of novel potential inhibitors will benefit from the use of these ligand-dependent protonation state patterns.

Conflicts of interest

There are no conflicts to declare.

Acknowledgements

JCG acknowledges support from the National Institutes of Health (R01-AI148740). JMP was supported by the DOE Office of Science through the National Virtual Biotechnology Laboratory, a consortium of DOE national laboratories focused on response to COVID-19, with funding provided by the Coronavirus CARES Act. This research used resources of the Oak Ridge Leadership Computing Facility at the Oak Ridge National Laboratory, which is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC05-00OR22725. This work also used resources, services, and support provided via the COVID-19 HPC Consortium (<https://covid19-hpc-consortium.org/>), which is a unique private-public effort to bring together government, industry, and academic leaders who are volunteering free compute time and resources in support of COVID-19 research. This work also used the Hive cluster, which is supported by the National Science Foundation under grant number 1828187 and is managed by the Partnership for an Advanced Computing Environment (PACE) at the Georgia Institute of Technology. This research

used resources at the Spallation Neutron Source and the High Flux Isotope Reactor, which are DOE Office of Science User Facilities operated by the Oak Ridge National Laboratory.

Notes and references

- 1 S. Ullrich and C. Nitsche, *Bioorg. Med. Chem. Lett.*, 2020, 30, 127377.
- 2 W. Dai, B. Zhang, H. Su, J. Li, Y. Zhao, X. Xie, Z. Jin, F. Liu, C. Li, Y. Li, F. Bai, H. Wang, X. Cheng, X. Cen, S. Hu, X. Yang, J. Wang, X. Liu, G. Xiao, H. Jiang, Z. Rao, L.-K. Zhang, Y. Xu, H. Yang and H. Liu, *Science*, 2020, 368, 1331–1335.
- 3 I. Schechter and A. Berger, *Biochem. Biophys. Res. Commun.*, 1967, 27, 157–162.
- 4 J. Anderson, C. Schiffer, S.-K. Lee and R. Swanstrom, *Antiviral Strat.*, 2009, 189, 85–110.
- 5 T. Pillaiyar, M. Manickam, V. Namasivayam, Y. Hayashi and S.-H. Jung, *J. Med. Chem.*, 2016, 59, 6595–6628.
- 6 L. Zhang, D. Lin, Y. Kusov, Y. Nian, Q. Ma, J. Wang, A. von Brunn, P. Leyssen, K. Lanko, J. Neyts, A. de Wilde, E. J. Snijder, H. Liu and R. Hilgenfeld, *J. Med. Chem.*, 2020, 63, 4562–4578.
- 7 L. Zhang, D. Lin, X. Sun, U. Curth, C. Drosten, L. Sauerhering, S. Becker, K. Rox and R. Hilgenfeld, *Science*, 2020, 368, 409–412.
- 8 Z. Jin, X. Du, Y. Xu, Y. Deng, M. Liu, Y. Zhao, B. Zhang, X. Li, L. Zhang, C. Peng, Y. Duan, J. Yu, L. Wang, K. Yang, F. Liu, R. Jiang, X. Yang, T. You, X. Liu, X. Yang, F. Bai, H. Liu, X. Liu, L. W. Guddat, W. Xu, G. Xiao, C. Qin, Z. Shi, H. Jiang, Z. Rao and H. Yang, *Nature*, 2020, 582, 289–293.
- 9 C. Ma, M. D. Sacco, B. Hurst, J. A. Townsend, Y. Hu, T. Szeto, X. Zhang, B. Tarbet, M. T. Marty, Y. Chen and J. Wang, *Cell Res.*, 2020, 30, 678–692.
- 10 X. Xue, H. Yu, H. Yang, F. Xue, Z. Wu, W. Shen, J. Li, Z. Zhou, Y. Ding, Q. Zhao, X. C. Zhang, M. Liao, M. Bartlam and Z. Rao, *J. Virol.*, 2008, 82, 2515–2527.
- 11 D. W. Kneller, G. Phillips, H. M. O'Neill, R. Jedrzejczak, L. Stols, P. Langan, A. Joachimiak, L. Coates and A. Kovalevsky, *Nat. Commun.*, 2020, 11, 3202.
- 12 W.-C. Hsu, H.-C. Chang, C.-Y. Chou, P.-J. Tsai, P.-I. Lin and G.-G. Chang, *J. Biol. Chem.*, 2005, 280, 22741–22748.
- 13 H. Chen, P. Wei, C. Huang, L. Tan, Y. Liu and L. Lai, *J. Biol. Chem.*, 2006, 281, 13894–13898.
- 14 H.-P. Chang, C.-Y. Chou and G.-G. Chang, *Biophys. J.*, 2007, 92, 1374–1383.
- 15 S.-C. Cheng, G.-G. Chang and C.-Y. Chou, *Biophys. J.*, 2010, 98, 1327–1336.
- 16 A. Douangamath, D. Fearon, P. Gehrtz, T. Krojer, P. Lukacik, C. D. Owen, E. Resnick, C. Strain-Damerell, A. Aimon, P. Ábrányi-Balogh, J. Brandaõ-Neto, A. Carbery, G. Davison, A. Dias, T. D. Downes, L. Dunnett, M. Fairhead, J. Firth, S. P. Jones, A. Keely, G. M. Kesserü, H. F. Klein, M. P. Martin, M. Noble, P. O'Brien, A. Powell, R. Reddi, R. Skyner, M. Snee, M. J. Waring, C. Wild, N. London, F. von Delft and W. M.A., 2020, *bioRxiv* 2020.05.27.118117.
- 17 H. Yang, M. Yang, Y. Ding, Y. Liu, Z. Lou, Z. Zhou, L. Sun, L. Mo, S. Ye, H. Pang, G. F. Gao, K. Anand, M. Bartlam, R. Hilgenfeld and Z. Rao, *PNAS*, 2003, 100, 13190–13195.
- 18 C. Huang, P. Wei, K. Fan, Y. Liu and L. Lai, *Biochemistry*, 2004, 43, 4568–4574.
- 19 S. Dajnowicz, R. C. Johnston, J. M. Parks, M. P. Blakeley, D. A. Keen, K. L. Weiss, O. Gerlits, A. Kovalevsky and T. C. Mueser, *Nat. Commun.*, 2017, 8, 955.
- 20 O. Gerlits, K. L. Weiss, M. P. Blakeley, G. Veglia, S. S. Taylor and A. Kovalevsky, *Sci. Adv.*, 2019, 5, eaav0482.
- 21 J. Tan, K. H. G. Verschuere, K. Anand, J. Shen, M. Yang, Y. Xu, Z. Rao, J. Bigalke, B. Heisen, J. R. Mesters, K. Chen, X. Shen, H. Jiang and R. Hilgenfeld, *J. Mol. Biol.*, 2005, 354, 25–40.
- 22 M. Macchiagodena, M. Pagliai and P. Procacci, *Chem. Phys. Lett.*, 2020, 750, 137489.
- 23 S. Kumar, P. P. Sharma, U. Shankar, D. Kumar, S. K. Joshi, L. Pena, R. Durvasula, A. Kumar, P. Kempaiah, Poonam and B. Rath, *J. Chem. Inf. Model.*, 2020, in press.
- 24 A. Fischer, M. Sellner, S. Neranjan, M. Smiesko and M. A. Lill, *Int. J. Mol. Sci.*, 2020, 21, 3626.
- 25 A. Aoudate, A. Ghaleb, S. Chtita, M. Aarjane, A. Ousaa, H. Maghat, A. Sbai, M. Choukrad, M. Bouachrine and T. Lakhli, *J. Biomol. Struct. Dyn.*, 2020, 1–14.
- 26 L. Mittal, A. Kumari, M. Srivastava, M. Singh and S. Asthana, *J. Biomol. Struct. Dyn.*, 2020, 1–19.
- 27 M. Bzówka, K. Mitusińska, A. Raczynska, A. Samol, J. A. Tuszyński and A. Góra, *Int. J. Mol. Sci.*, 2020, 21, 3099.
- 28 D. Suarez and N. Diaz, *J. Chem. Inf. Model.*, 2020, in press.
- 29 A. Poater, *J. Phys. Chem. Lett.*, 2020, 6262–6265.
- 30 A. Wlodawer and L. Sjölin, *Proc. Natl. Acad. Sci. U.S.A.*, 1981, 78, 2853–2855.
- 31 A. Y. Kovalevsky, T. Chatake, N. Shibayama, S. Y. Park, T. Ishikawa, M. Mustyakimov, Z. Fisher, P. Langan and Y. Morimoto, *J. Mol. Biol.*, 2010, 398, 276–291.
- 32 P. Kumar, E. H. Serpersu and M. J. Cuneo, *Sci. Adv.*, 2018, 4, eaas8667.
- 33 D. W. Kneller, G. Phillips, K. L. Weiss, S. Pant, Q. Zhang, H. M. O'Neill, L. Coates and A. Kovalevsky, *J. Biol. Chem.*, 2020, In press.
- 34 M. Lazaratos, K. Karathanou and A.-N. Bondar, *Curr. Opin. Struct. Biol.*, 2020, 64, 79–87.
- 35 K. Karathanou, M. Lazaratos, É. Bertalan, M. Siemers, K. Buzar, G. F. Schertler, C. del Val and A.-N. Bondar, 2020, *bioRxiv* 2020.06.23.164947.
- 36 A. Paasche, A. Zipper, S. Sch fer, J. Ziebuhr, T. Schirmeister and B. Engels, *Biochemistry*, 2014, 53, 5930–5946.
- 37 J. Yin, C. Niu, M. M. Cherney, J. Zhang, C. Huitema, L. D. Eltis, J. C. Vederas and M. N. G. James, *J. Mol. Biol.*, 2007, 371, 1060–1074.
- 38 K. ?widerek and V. Moliner, *Chem. Sci.*, 2020, 11, 10626–10630.

- 39 W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey and M. L. Klein, *J. Chem. Phys.*, 1983, 79, 926–935.
- 40 J. Huang, S. Rauscher, G. Nawrocki, T. Ran, M. Feig, B. L. de Groot, H. Grubmüller and A. D. MacKerell, *Nat. Methods*, 2017, 14, 71–73.
- 41 K. Vanommeslaeghe, E. Hatcher, C. Acharya, S. Kundu, S. Zhong, J. Shim, E. Darian, O. Guvench, P. Lopes, I. Vorobyov and A. D. MacKerell Jr., *J. Comput. Chem.*, 2010, 31, 671–690.
- 42 K. Vanommeslaeghe and A. D. MacKerell, *J. Chem. Inf. Model.*, 2012, 52, 3144–3154.
- 43 K. Vanommeslaeghe, E. P. Raman and A. D. MacKerell, *J. Chem. Inf. Model.*, 2012, 52, 3155–3168.
- 44 C. G. Mayne, J. Saam, K. Schulten, E. Tajkhorshid and J. C. Gumbart, *J. Comput. Chem.*, 2013, 34, 2757–2770.
- 45 M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, G. Scalmani, V. Barone, G. A. Petersson, H. Nakatsuji, X. Li, M. Caricato, A. V. Marenich, J. Bloino, B. G. Janesko, R. Gomperts, B. Menucci, H. P. Hratchian, J. V. Ortiz, A. F. Izmaylov, J. L. Sonnenberg, D. Williams-Young, F. Ding, F. Lipparini, F. Egidi, J. Goings, B. Peng, A. Petrone, T. Henderson, D. Ranasinghe, V. G. Zakrzewski, J. Gao, N. Rega, G. Zheng, W. Liang, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, T. Vreven, K. Throssell, J. A. Montgomery, Jr., J. E. Peralta, F. Ogliaro, M. J. Bearpark, J. J. Heyd, E. N. Brothers, K. N. Kudin, V. N. Staroverov, T. A. Keith, R. Kobayashi, J. Normand, K. Raghavachari, A. P. Rendell, J. C. Burant, S. S. Iyengar, J. Tomasi, M. Cossi, J. M. Millam, M. Klene, C. Adamo, R. Cammi, J. W. Ochterski, R. L. Martin, K. Morokuma, O. Farkas, J. B. Foresman and D. J. Fox, *Gaussian 16 Revision B.01*, 2016, Gaussian Inc. Wallingford CT.
- 46 J. C. Phillips, R. Braun, W. Wang, J. Gumbart, E. Tajkhorshid, E. Villa, C. Chipot, R. D. Skeel, L. Kale and K. Schulten, *J. Comp. Chem.*, 2005, 26, 1781–1802.
- 47 J. C. Phillips, D. J. Hardy, J. D. C. Maia, J. E. Stone, J. V. Ribeiro, R. C. Bernardi, R. Buch, G. Fiorin, J. Hénin, W. Jiang, R. McGreevy, M. C. R. Melo, B. K. Radak, R. D. Skeel, A. Singharoy, Y. Wang, B. Roux, A. Aksimentiev, Z. Luthey-Schulten, L. V. Kalé, K. Schulten, C. Chipot and E. Tajkhorshid, *J. Chem. Phys.*, 2020, 153, 044130.
- 48 T. A. Darden, D. M. York and L. G. Pedersen, *J. Chem. Phys.*, 1993, 98, 10089–10092.
- 49 S. E. Feller, Y. H. Zhang, R. W. Pastor and B. R. Brooks, *J. Chem. Phys.*, 1995, 103, 4613–4621.
- 50 C. Owen, P. Lukacik, C. Strain-Damerell, A. Douangamath, A. Powell, D. Fearon, J. Brandao-Neto, A. Crawshaw, D. Aragao, M. Williams, R. Flaig, D. Hall, K. McAuley, M. Mazzorana, D. Stuart, F. von Delft and M. Walsh, SARS-CoV-2 main protease with unliganded active site (2019-nCoV, coronavirus disease 2019, COVID-19), 2020, RCSB Protein Data Bank (PDB) ID 6YB7 3-7.
- 51 W. Humphrey, A. Dalke and K. Schulten, *J. Mol. Graph.*, 1996, 14, 33–38.
- 52 M. J. Abraham, T. Murtola, R. Schulz, S. Páll, J. C. Smith, B. Hess and E. Lindahl, *SoftwareX*, 2015, 1-2, 19–25.
- 53 M. Parrinello and A. Rahman, *J. Appl. Phys.*, 1981, 52, 7182–7190.
- 54 B. Hess, *J. Chem. Theory Comput.*, 2008, 4, 116–122.
- 55 L. D. Landau, *Statistical physics*, The Clarendon Press, Oxford, 1938.
- 56 R. W. Zwanzig, *J. Chem. Phys.*, 1954, 22, 1420–1426.
- 57 J. Gao, K. Kuczera, B. Tidor and M. Karplus, *Science*, 1989, 244, 1069–1072.
- 58 A. Pohorille, C. Jarzynski and C. Chipot, *J. Phys. Chem. B*, 2010, 114, 10235–10253.
- 59 C. H. Bennett, *J. Comput. Phys.*, 1976, 22, 245–268.
- 60 H. Chen, J. D. C. Maia, B. K. Radak, D. J. Hardy, W. Cai, C. Chipot and E. Tajkhorshid, *J. Chem. Inf. Model.*, 2020.
- 61 B. Laurent, M. Chavent, T. Cragnolini, A. C. E. Dahl, S. Pasquali, P. Derreumaux, M. S. Sansom and M. Baaden, *Bioinformatics*, 2015, 31, 1478–1480.
- 62 R. Hilgenfeld, *FEBS J.*, 2014, 281, 4085–4096.
- 63 R. Anandakrishnan, B. Aguilar and A. V. Onufriev, *Nucleic Acids Res.*, 2012, 40, W537–541.
- 64 M. H. Olsson, C. R. Søndergaard, M. Rostkowski and J. H. Jensen, *J. Chem. Theory Comput.*, 2011, 7, 525–537.
- 65 M. O. Kim, S. E. Nichols, Y. Wang and J. A. McCammon, *J. Comput. Aided Mol. Des.*, 2013, 27, 235–246.
- 66 Y. Huang, W. Chen, J. A. Wallace and J. Shen, *J. Chem. Theory Comput.*, 2016, 12, 5411–5421.
- 67 J. Henderson, N. Verma and S. J., 2020, bioRxiv 2020.06.30.181305v1.
- 68 S. Scheiner, T. Kar and J. Pattanayak, *J. Am. Chem. Soc.*, 2002, 124, 13257–13264.
- 69 P. Zhou, F. Tian, F. Lv and Z. Shang, *Proteins*, 2009, 76, 151–163.
- 70 C.-P. Chuck, L.-T. Chong, C. Chen, H.-F. Chow, D. C.-C. Wan and K.-B. Wong, *PLoS One*, 2010, 5, e13197.
- 71 A. S. Hauser, M. M. Attwood, M. Rask-Andersen, H. B. Schiöth and D. E. Gloriam, *Nat. Rev. Drug Discovery*, 2017, 16, 829–842.
- 72 F. Deflorian, L. Perez-Benito, E. B. Lenselink, M. Congreve, H. W. T. van Vlijmen, J. S. Mason, C. de Graaf and G. Tresadern, *J. Chem. Inf. Model.*, 2020, doi: 10.1021/acs.jcim.0c00449.
- 73 G. Madhavi Sastry, M. Adzhigirey, T. Day, R. Annabhimoju and W. Sherman, *J. Comput. Aided Mol. Des.*, 2013, 27, 221–234.
- 74 E. Awoonor-Williams, A. G. Walsh and C. N. Rowley, *Biochim. Biophys. Acta Proteins Proteom.*, 2017, 1865, 1664–1675.