Unraveling the SARS-CoV-2 Main Protease Mechanism Using Multiscale DFT/MM Methods

Carlos A. Ramos-Guzmán, J. Javier Ruiz-Pernía*, Iñaki Tuñón*

Departamento de Química Física, Universidad de Valencia, 46100 Burjassot (Spain)

*To whom correspondence should be addressed:
ignacio.tunon@uv.es
j.javier.ruiz@uv.es

Keywords: 3CL protease, SARS-CoV-2, Minimum Free Energy Path, QM/MM, Acylation, Deacylation, Molecular Dynamics

Abstract

We present a detailed theoretical analysis of the reaction mechanism of proteolysis catalyzed by the main protease of SARS-CoV-2. Using multiscale simulation methods, we have characterized the interactions stablished by a peptidic substrate in the active site and then we have explored the free energy landscape associated to the acylation and de-acylation steps of the proteolysis reaction, characterizing the transition states of the process. Our mechanistic proposals can explain most of the experimental observations made on the highly similar ortholog protease of SARS-CoV. We point out to some key interactions that may facilitate the acylation process and thus can be crucial in the design of more specific and efficient inhibitors of the main protease activity. In particular, from our results, the P1' residue can be a key factor to improve the thermodynamics and kinetics of the inhibition process.

40

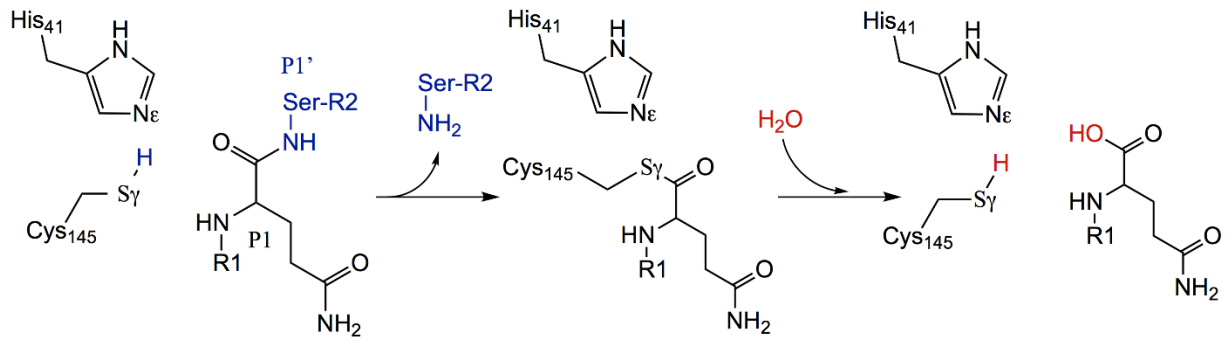
41 The recent outbreak of COVID-19, a pneumonia-like illness caused by a coronavirus named as
42 SARS-CoV-2[1], has rapidly evolved into a pandemic as recognized by the World Health
43 Organization. SARS-CoV-2 has been shown to be highly contagious, causing a large number of
44 infections around the world. The absence of vaccines and specific treatments has contributed to
45 a rapid spread of this disease and a fatal outcome in many cases. Furthermore, the existence of
46 other similar virus detected in animals opens the possibility of future similar diseases.[2,3] Thus,
47 finding effective strategies for the identification of potential targets for new drugs to fight against
48 SARS-CoV-2 and other CoV-like virus is, nowadays, an urgent need. One of these strategies is
49 based in the disruption of the activity of those enzymes that are crucial in the replication cycle of
50 the virus using adequate compounds. In this sense, the knowledge of the catalytic activity of the
51 enzyme at atomistic detail is one of the more powerful tools for efficient and specific new drugs
52 design. In particular, the characterization and analysis of the geometry and electronic properties
53 of the reaction Transition State (TS), can be used as a guide for the design of active site inhibitors.
54 In this work we analyze the reaction mechanisms for the main protease of SARS-CoV-2, also
55 referred to as 3C-like protease ($3CL^{pro}$) using DFT-based multiscale methods. This enzyme plays
56 an essential role during the replication of the virus and has not closely related homologues in
57 human beings, making it an attractive drug target[4].

58

59 The $3CL^{pro}$ enzyme of SARS-CoV-2 exists as a functional homodimer with two active sites in
60 charge of cleaving the translated polyproteins into individual fragments to be used by the
61 coronavirus[5]. As other cysteine proteases, each of the active sites contains a Cys-His catalytic
62 dyad in charge of the hydrolysis of the peptide bond at specific sites of a polypeptide chain.
63 Several structures of this protease have been already resolved by means of x-ray crystallography
64 and deposited in the Protein Data Bank (PDB), including the free protease (PDB codes 6Y2E[2]
65 and 6Y84[6]) and inhibitor bound proteases (PDB codes 6LU7[6], 6Y2F[7] and 6LZE[8]). The SARS-
66 CoV-2 main protease has a structure virtually identical to the ortholog from SARS-CoV (96%
67 identity). Even more, the main residues involved in catalysis, binding and dimerization processes
68 are fully conserved[9]. Consequently, these two ortholog enzymes display highly similar substrate
69 preferences[10]. The substrate cleavage by the $3CL^{pro}$ takes place between Gln at the P1 position
70 and a Gly/Ala/Ser at the P1' one (P and P' identify the residues placed before and after the scissile
71 bond, respectively), being the presence of Gln an essential requirement[11].

72

In principle, the reaction mechanism of cysteine proteases involves two basic steps (see Scheme 1)[12]. In the first step, acylation, the peptide bond is broken, releasing the P' fragment of the peptidic substrate and forming an acyl-enzyme complex where the catalytic cysteine (Cys145 in the protease of SARS-CoV-2) is covalently bound to the carbon atom of the P1 residue of the target peptide. In a second step, de-acylation, the acyl-enzyme is hydrolyzed, releasing the P fragment and recovering the enzymatic active site for another catalytic cycle. Covalent inhibitors of the protease activity form acyl-enzyme complexes that cannot be hydrolyzed, remaining bonded into the active site[13,14]. In this work we take benefit of the similarities between the proteases of SARS-CoV and SARS- CoV-2 viruses and the existence of ligand-bound structures to build a structural model of a peptide substrate-enzyme Michaelis complex and to study the reaction mechanism using computational simulations.



Scheme 1

Results

As detailed in the Methods section we carried out Molecular Dynamics (MD) simulations (3 replicas) of the peptide substrate-protease complex built from the 6Y2F PDB structure[7]. The chosen substrate for our simulation, Ac-Ser-Ala-Val-Leu-Gln-Ser-Gly-Phe-NMe, was selected from a structure of the SARS-CoV protease structure. For this substrate, the proteolysis takes place in the Gln(P1)-Ser(P1') bond. A total of 10 μs of classical simulations were run using the AMBER19 GPU version of *pmemd*[15,16]. We then explored the reaction mechanism using multiscale simulation methods at the B3LYPD3/MM level, with the 6-31+G* basis set, as explained in Methods section. The string-method[17,18] was employed to find the minimum free energy paths (MFEP) on multidimensional free energy surfaces and to trace the associated free energy profiles.

99

100 ***Enzyme-Substrate Complex***

101 The time evolution of the root mean square deviation (RMSD) values for each replica (2 protomers
102 and 2 substrates in each replica) are shown in Figure S1. These values show that the protein
103 structure is well-equilibrated and there are no large differences with respect to the initial
104 structures, prepared from the x-ray data. The observations made on the three replicas (one of 8
105 μs and two of 1 μs) are very similar in all cases. The substrate-binding pocket is divided into a
106 series of subsites (denoted as S and S'), each accommodating a single residue of the substrate
107 placed before (P) or after (P') the scissile peptide bond. The map of hydrogen bond interactions
108 observed during our MD simulations of the Michaelis complex is given in Figure 1a, a general
109 view of the substrate in the two active sites of the dimer formed by protomers A and B is shown
110 in Figure 1b, while an insight into the active site of protomer A is provided in Figure 1c. The $3CL^{pro}$
111 of SARS-CoV-2 (as is also the case of the SARS-CoV ortholog) presents a high specificity for Gln
112 at P1 position[11,19]. As seen in Figure 1a the P1 residue is the one establishing more hydrogen
113 bond interactions with the enzyme. The O and N main chain atoms of Gln-P1 are found to make
114 hydrogen bonds with main chain atoms of Gly143, Ser144 and His164 (unless indicated all the
115 residues of the protein belong to the same protomer A). Regarding the side chain of Gln-P1, this
116 is accommodated into the S1 subsite through hydrogen bond contacts with the main chain atoms
117 of Phe140 and Leu141, with the Nε atom of His163 and the Oε atoms of Glu166. This last residue
118 is in turn hydrogen bonded to the terminal NH group of Ser1 from protomer B. In fact, the N-
119 terminal fragment (N-finger) of protomer B plays an active role pre-organizing the active site of
120 protomer A for catalysis[20]. Dimerization is an essential condition for catalysis in the protease of
121 related coronavirus[19,21,22] and, consequently, those mutants lacking the N-finger fragment are
122 almost completely inactive[23]. The side chain of Leu-P2 is surrounded by the side chains of His41,
123 Met49, His164, Met165 and Asp187, while the main chain amide group is hydrogen bonded to
124 the Oε atom of Gln189. The side chain of Val-P3 is solvent-exposed, while main chain N and O
125 atoms are hydrogen bonded to main chain atoms of Glu166.

126

127 The binding subsite for Ala-P4 is constituted by main chain interactions with Gln189 and Thr190,
128 while the side chain of this residue is placed between the side chains of Met165, Leu167 and
129 Pro168. The S5 subsite is formed by the side chain of Pro168 and by main chain atoms of Thr190.
130 This description of the S1-S5 interaction subsites that have been observed in our simulations

agrees with the descriptions found in the x-ray structures of SARS-CoV-2 3CL protease with inhibitor bounds in the active site[8,13,14]. Our MD simulations of the substrate-enzyme complex offers, in addition, a detailed description of the S' subsites, those that accommodate the P' residues placed after the scissile peptide bond. The main chain O atom of Ser-P1' establishes a hydrogen bond with the amide group of Gly143 and the side chain of Asn142. The hydroxyl group of the P1' side chain can contact with the catalytic dyad (Cys145 and His41) while the CH_2 group is packed between Thr25, Thr26 and Leu27. Gly-P2' is stabilized through main chain contacts with Thr25 and Thr26. Finally, the side chain of Phe-P3' residue is packed against the side chain of Thr24. This structural information can be useful in order to improve the binding and specificity of potential inhibitors of the protease activity because these structural findings are lost in the x-ray structures obtained from those inhibitors in which the fragment corresponding to P' residues either is released during the formation of the acyl-enzyme complex[14] or is smaller than in our substrate[8,13].

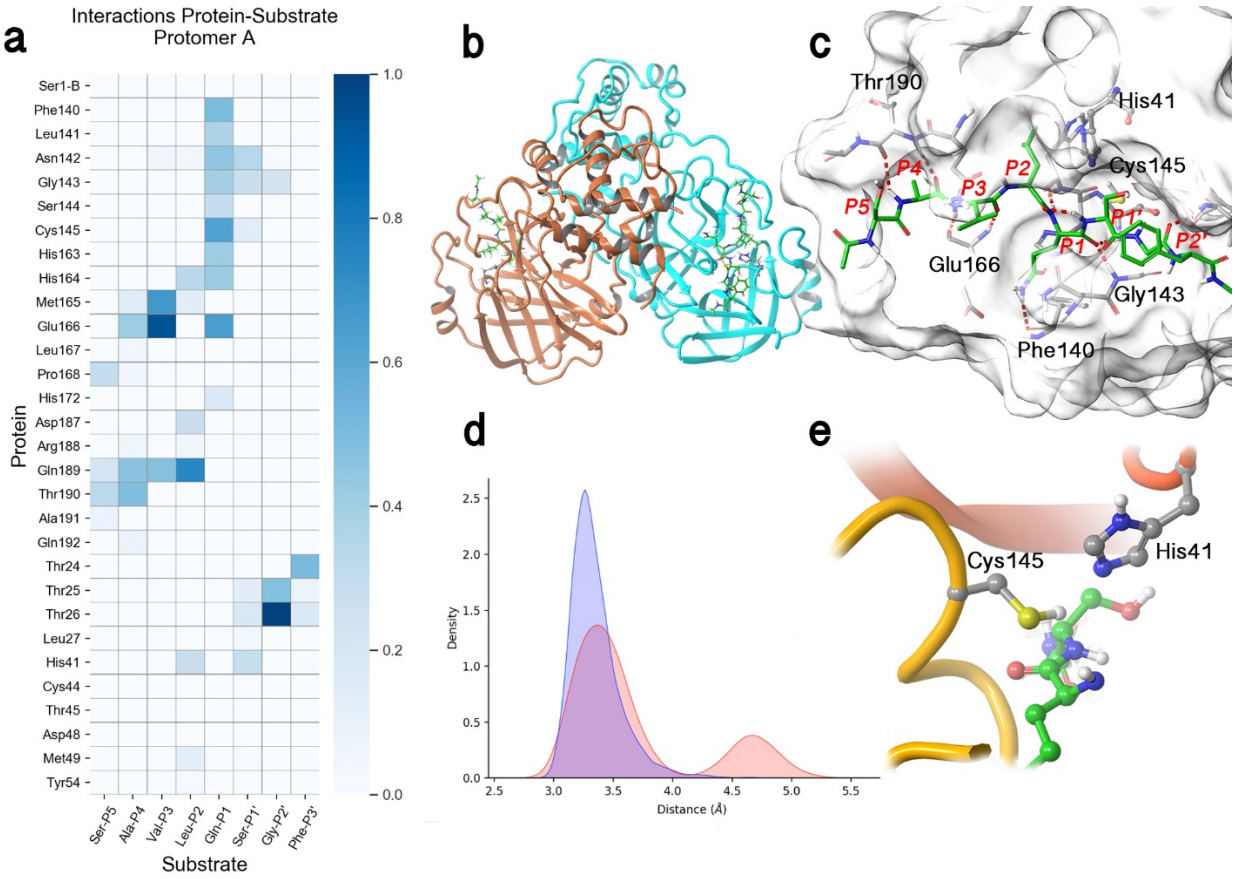


Figure 1. Results of the Molecular Dynamics simulation of 3CL^{pro} of SARS-CoV-2 in complex with the substrate with sequence Ac-Ser-Ala-Val-Leu-Gln-Ser-Gly-Phe-NMe. (**1a**) Fraction of hydrogen bond contacts between the residues of the substrate and those of the protease found during the trajectory of the Michaelis Complex. A hydrogen bond contact is counted when the donor-acceptor distance is < 3.8 Å and the hydrogen bond angle is > 120º. (**1b**) General overview of the substrate-enzyme complex, showing the dimeric nature of the protease with two identical active sites occupied by the substrate. Note that the N-finger of each protomer is close to the active site of the neighbor protomer. (**1c**) Insight into the binding pose of the peptide substrate into the active site, showing the most important active site residues and the positions occupied by the P and P' residues of the substrate. (**1d**) Probability densities of the distances from the Cys145-Sγ atom to the carbonyl carbon atom of the substrate (C(P1)), in red, and to the Nε atom of His41, in blue. The bimodal distribution of Sγ-C(P1) distances correspond to the *trans* (shorter distances) and *gauche* (longer distances) conformations of Cys145. (**1e**) Disposition of the substrate in the vicinity of the catalytic dyad when Cys145 is present in the *trans* conformation. Note the proximity between Sγ and C(P1) atoms and the orientation of the sulfhydryl proton towards the Nε atom of His41.

165 ***The Catalytic Dyad***

166 The reaction mechanism of cysteine proteases involves the nucleophilic attack of the Sγ atom of
167 a cysteine (Cys145 in our case) to the C(P1) atom of the peptide bond. Figure 1d shows the
168 probability distribution of Sγ-C(P1) distances found during our MD simulations. The distribution is
169 clearly bimodal, with two peaks centered at 3.4 and 4.7 Å. These two peaks correspond to two
170 different conformations of the side chain of Cys145, which can be present in *trans* and *gauche*
171 conformations. This is in agreement with the observations made on the x-ray structure of the
172 orthologue protease of SARS-CoV[20]. The most probable conformation corresponds to the *trans*
173 conformer in which the Sγ sulfur atom is closer to the substrate (see Figure 1e). In both
174 conformations the catalytic dyad remains hydrogen bonded, being the most probable distance
175 between Cys145-Sγ and His41-Nε of about 3.3 Å (see Figure 1d). Interestingly, His41 is, in turn,
176 hydrogen bonded, through a highly conserved crystallographic water molecule, to Asp187. This
177 interaction can raise the pK_a of the histidine, increasing its ability to work as a base and abstract
178 the proton from Cys145.
179
180 Considering the short distance observed between Cys145 and His41 and the possible activation
181 of this last residue by the nearby Asp187, we explored the possibility to find the catalytic dyad
182 forming an ion pair ($Cys^-/HisH^+$ or IP in Figure 2) instead of the neutral form modelled in the
183 Michaelis complex (CysH/His or MC in Figure 2). We evaluated the free energy difference
184 between these two forms of the dyad by means of free energy profiles associated to the proton
185 transfer coordinate from Cys145 to His41 obtained at the B3LYPD3/MM level (see Methods). The
186 proton transfer free energy profiles (Figure 2a) were obtained for the holo and apo forms of $3CL^{pro}$.
187 According to Figure 2a the catalytic dyad is more stable in its neutral form, both for the apo and
188 holo forms. The IP is 2.9 and 4.8 $kcal \cdot mol^{-1}$ above the neutral form, in the apo and holo enzymes,
189 respectively. The anionic Cys145 can be stabilized by the presence of water molecules and by
190 the hydroxyl group of Ser-P1' (see Figure 2b). Bulkier residues at the P1' position could hinder
191 the access of water molecules and thus destabilize the unprotonated form of Cys145. This could
192 be one of the factors explaining the preference of $3CL^{pro}$ for small residues at the P1' position
193 (Ser/Ala/Gly). According to our free energy profiles shown in Figure 2a the IP form is better
194 stabilized with respect to the neutral dyad in the apo enzyme than in the holo one (by almost 2
195 $kcal \cdot mol^{-1}$), which can be related to the better solvation of the anionic cysteine in the former. In
196 both the holo and apo enzymes, the free energy barrier associated to the transfer of the proton

between His41 to Cys145 is small, revealing a fast equilibrium between the ion pair and neutral versions of the dyad, being the latter the predominant form. The existence of a low-lying IP dyad is compatible with the experimental observations made in the kinetic characterization of the highly homologue 3CL^{pro} of SARS-CoV, in which an ion pair mechanism for the proteolysis was proposed on the basis of the pH-inactivation profile with iodoacetamide and the analysis of solvent isotope effects[24].

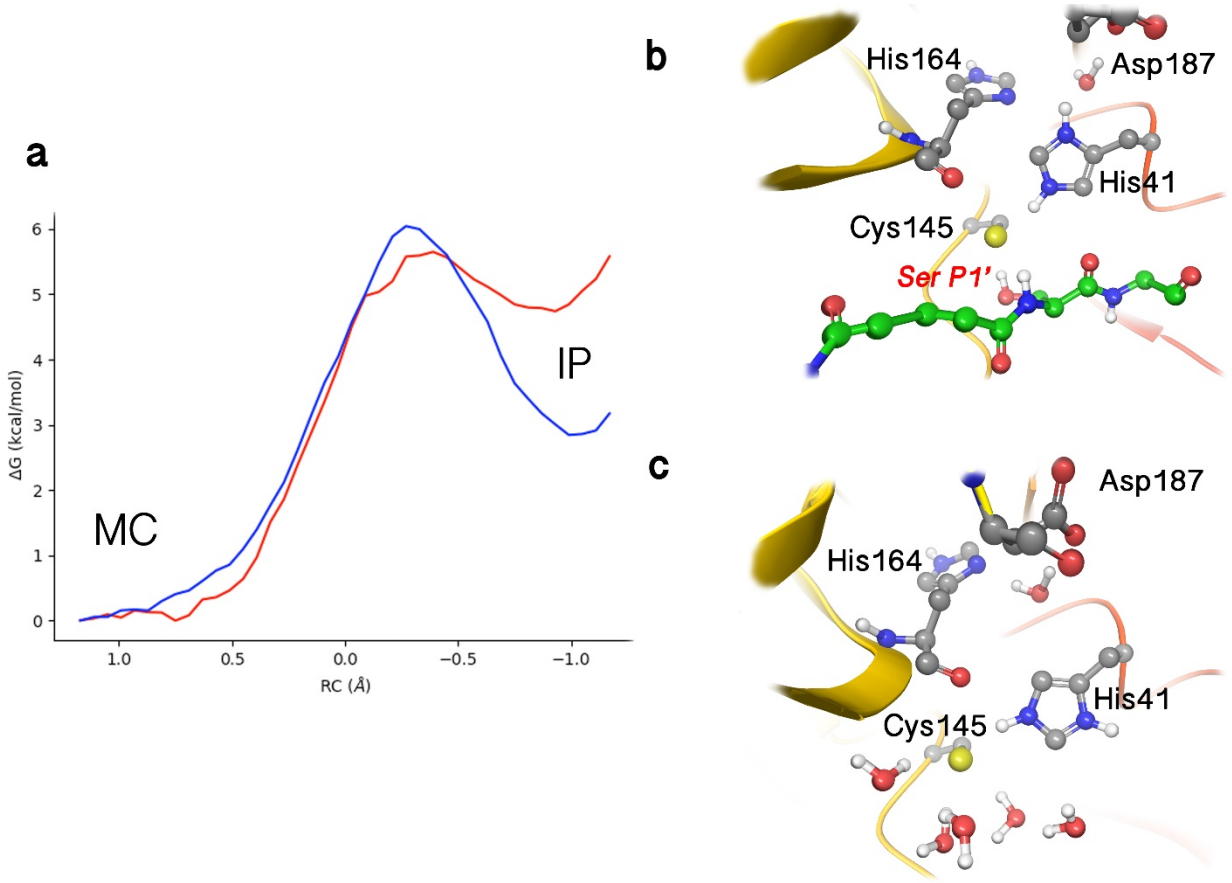


Figure 2. Analysis of the formation of an ion pair (IP) catalytic dyad (Cys145⁻···His41H⁺) from the neutral form (Cys145H····His41) found in the Michaelis Complex (MC). (**2a**) B3LYPD3/6-31+G*/MM free energy profile associated to the proton transfer coordinate from the Sγ atom of Cys145 to the Nε atom of His41 (d(Nε-H)-d(Sγ-H)) in the apo (blue line) and holo (red line) enzymes. (**2b**) Representation of the ion pair in the holo enzyme, showing those interactions stabilizing the charged states of the catalytic dyad, in particular the hydroxyl group of Ser(P1'). (**2c**) Representation of the ion pair in the apo enzyme, showing the presence of water molecules stabilizing the charged catalytic dyad when the substrate is absent.

216
217
218 ***The Acylation step***
219 We explored the free energy landscape for the formation of the acyl-enzyme starting from the
220 catalytic dyad IP at the B3LYPD3/MM level. The converged MFEP is shown in Figures 3a-b.
221 According to our simulations, after ion pair formation, the acylation proceeds by means of a proton
222 transfer from His41 to the N(P1') atom followed by the nucleophilic attack of Cys145-Sγ on the
223 C(P1) atom and the simultaneous breaking of the C(P1)-N(P1') peptide bond. These elementary
224 events take place in a concerted but asynchronous way. The transition state (TS) found for this
225 mechanism (see Figure 3c) is associated to the proton transfer from His41 to the amide nitrogen
226 atom of the peptide bond N(P1'). At the TS the Sγ atom of Cys145 approaches to the C(P1) atom,
227 reducing the interatomic distance from 3.11 to 2.34 Å. This approach is accompanied by a
228 moderate lengthening of the peptide bond (the C(P1)-N(P1') distance being lengthened from 1.36
229 to 1.54 Å). According to the free energy path shown in Figure 3a, the total free energy barrier
230 associated to the acylation process, including the free energy cost of the ion pair formation, is of
231 14.6 kcal·mol^{-1}. This value is compatible with the activation free energies derived from the steady-
232 state rate constants measured at 25º C for peptides cleaved at the Gln-Ser bond by the highly
233 similar ortholog 3CL^{pro} of SARS-CoV (between 16.2 and 17.2 kcal·mol^{-1})[24]. It must be noticed that
234 in this proteolysis the acylation step is not considered to be the rate-limiting one and then these
235 experimental activation free energies provide an upper limit for the acylation barrier[24]. Figure 3c
236 shows that this TS is stabilized by means of a hydrogen bond interaction with the hydroxyl group
237 of SerP1', indicating, also in agreement with the experimental observations, that the leaving group
238 plays an important role in catalysis. Remarkably, the proposed reaction mechanism is also in
239 good agreement with the experimental proton inventory results, that indicate that there are two
240 protons in flight during the acylation, one at the TS and another one at earlier stages[24]. In our
241 picture these two proton transfers events correspond to the proton transferred from His41 to the
242 N(P1') atom of the substrate at the TS and the proton transfer from Cys145 to His41 during IP
243 formation. Regarding the acyl-enzyme product (Figure 3d), our free energy profile shows two
244 possible conformations that differ in the presence of a water molecule that plays a key role during
245 de-acylation (see Figure S2). Finally, in our free energy profile the formation of the acyl-enzyme
246 (see Figure 3d) is almost thermo-neutral, with a reaction free energy of about -1 kcal·mol^{-1}.
247

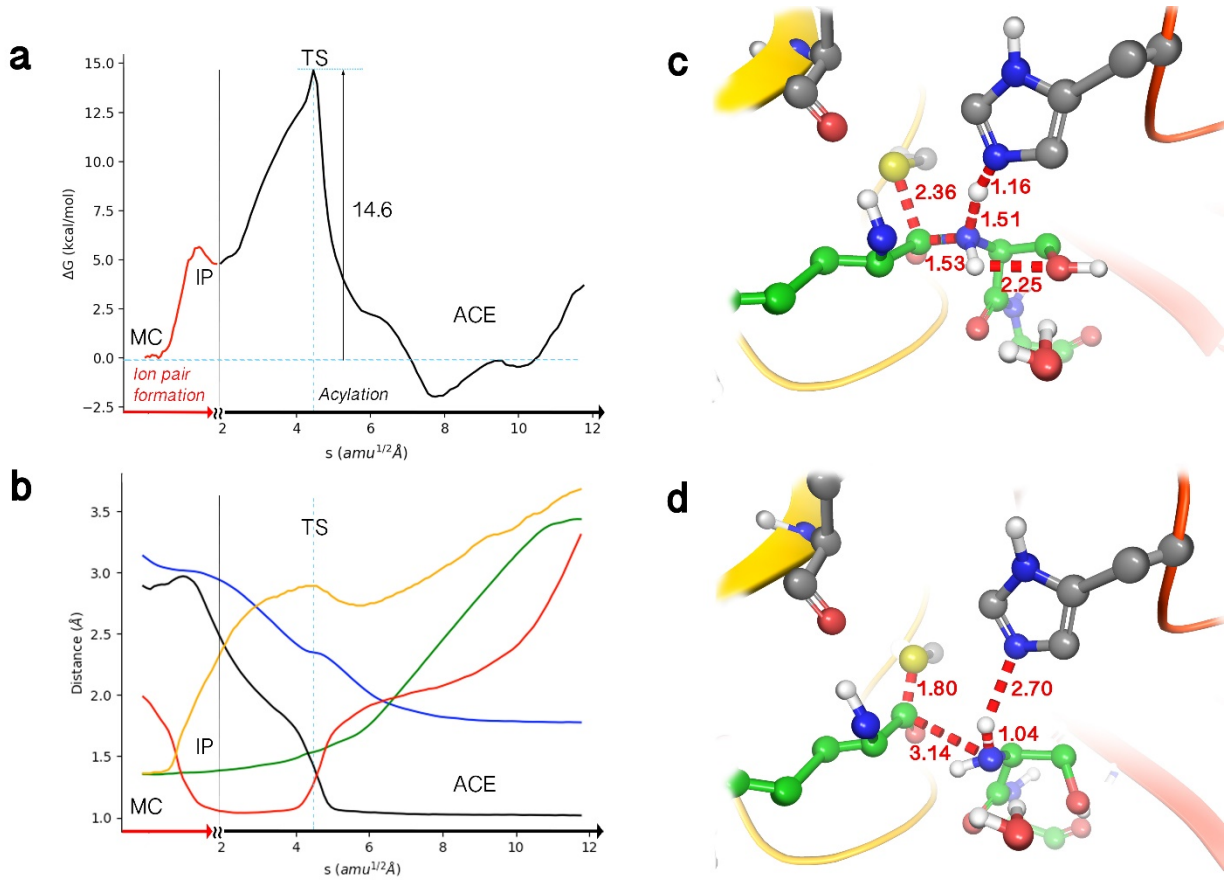


Figure 3. Simulation of the acylation reaction taking place through the formation of the ion pair. (**3a**) B3LYPD3/6-31+G*/MM free energy profile along the path-CV for the acylation reaction after the formation of the ion pair (IP) from the Michaelis complex (MC). The reaction takes place with a single transition state (TS) that yield the acyl-enzyme (ACE), which can be present in two conformations. (**3b**) Evolution of the distances selected as Collective Variables (CVs) along the Minimum Free Energy Path (MFEP). Sγ-H in yellow, H-Nε in red, C(P1)-N(P1') (the scissile peptide bond) in green, H-N(P1') in black, Sγ-C(P1) in blue. (**3c**) Representation of the TS for the acylation process. This TS corresponds to the proton transfer from His41 in the ion pair catalytic dyad to the nitrogen atom of peptide bond (N(P1') with the approach of the Sγ atom of Cys145 to the carbonyl carbon atom (C(P1)) and the lengthening of the peptide bond. The values of the distances correspond to the coordinates of the MFEP at the TS, except the intramolecular distance between the hydroxyl and NH group of Ser(P1') that has been averaged over the trajectory of the corresponding string node. (**3d**) Representation of the acyl-enzyme complex formed between the enzyme and the P fragment of the peptide, with a water molecule hydrogen bonded to the N-terminal group of the P' fragment. The free energy profile shows two minima for the acyl-enzyme complex, differing in the distance between the P and P' fragments, as shown in Figure S2.

Very recently an interesting QM/MM study of the same protease with a different substrate (where the leaving group was not a peptide fragment but a fluorescent tag, 7-amino-4-carbamoylmethylcoumarin) has been reported[25]. In that work the proton transfer from Cys145 to His41 was found to be concomitant with the nucleophilic attack of the Sγ atom on the carbonyl carbon atom, forming a thiohemiketal intermediate and the cleavage of the peptide bond takes place in a subsequent step assisted by the proton transfer from His41. The differences with respect to our results could be due to the use of a non-peptidic leaving group and/or the use of different theoretical descriptions (that work used a combination of semiempirical and DFT methods with the M06-2X functional). In any case, that study obtained an activation free energy of 19.9 kcal·mol^{-1}, in excellent agreement (within 0.5 kcal·mol^{-1}) with the value derived from the experimental rate constant obtained for that substrate[10]. Interestingly, this rate constant (0.050 s^{-1})[10] is significantly smaller than the value reported for the hydrolysis of the Gln-Ser bond by the ortholog enzyme of SARS-CoV (1.5-8.5 s^{-1})[24]. The gap between these two experimental rate constant values could be due to differences in the preparation and purification of the enzyme or to genuine mechanistic differences between substrates in the main protease[24,26]. In this sense, as discussed above, the presence of a hydroxyl group at P1' position could play an important role in the acylation process, improving the binding and kinetics of a hypothetical inhibitor.

In order to explore other possible mechanisms[27] we also studied a reaction path that does not involve the formation of an ion pair, although the associated free energy barrier is considerably higher and incompatible with the experimental rate constant (see Figure S3).

The De-acylation step

Regarding the de-acylation step, the standard mechanistic proposal suggested for related enzymes[28] assumes that once the neutral P'-NH_2 peptide fragment has left the active site, a water molecule activated by His41 attacks the C(P1)-Sγ bond, releasing the P-COOH peptide (also with a neutral terminal group) and, finally, regenerating the enzyme after a proton transfer from His41 to Cys145 (see Scheme 1). In our simulations of the acylation product we found that a water molecule can be placed in between the P'-NH_2 leaving fragment and the acyl-enzyme complex, being correctly oriented to perform the hydrolysis of the acyl-enzyme (see Figure 3d and Figure S2). This configuration suggests an alternative reaction mechanism that can yield the two peptide

300 fragments with correct protonation states in the terminal groups and regenerate the enzymatic
301 active site in its most stable state (the neutral catalytic dyad). An additional advantage of this
302 novel mechanism is that involves water activation by means of the N-terminus of the P' fragment,
303 which is known to be a better base than histidine side chains (the average pK_a values are about
304 7.7 and 6.6, respectively)[29]. Finally, the proposed mechanism can be at the origin of the
305 differences observed when the scissile bond is an amide instead of an ester (in that case a basic
306 N-terminal group is not formed after the acylation).[24] We obtained the MFEP corresponding to
307 such a mechanism at the B3LYPD3/MM level (see Figure 4). This process is stepwise, presenting
308 two TSs (see Figure 4a). The first TS (TS1 in Figure 4c) corresponds to the proton transfer from
309 the water molecule to the N(P1') atom, resulting in the formation of the P'-NH_3^+ peptide fragment.
310 The free energy barrier associated to this step is of 15.6 kcal·mol^{-1}, in excellent agreement with
311 the values derived from the reaction rate constants for the ortholog protease of SARS-CoV (from
312 16.2 to 17.2 kcal·mol^{-1}).[24] This proton transfer is concomitant to the attack of the hydroxyl group
313 on the C(P1) carbonyl carbon atom, resulting in the formation of an intermediate thiodiolate
314 (Figure 4d). After rotation of the hydroxyl group to orient the proton towards the sulfur atom, the
315 reaction proceeds with the cleavage of the C(P1)-Sγ bond. The second TS observed during the
316 de-acylation (TS2, Figure 4e), corresponds to the separation of the Sγ atom (the C(P1)-Sγ
317 distance being 2.67 Å). The free energy barrier associated to this second step from the acyl-
318 enzyme complex is very close to the first one, 15.8 kcal·mol^{-1}. Afterwards, the leaving cysteine is
319 stabilized by means of a proton transfer from the C-terminal group to the Sγ atom, regenerating
320 the enzyme in its more stable protonation state (a neutral catalytic dyad) and yielding the P
321 peptide fragment with a terminal unprotonated carboxylate (the product is represented in Figure
322 4f). The proposed mechanism, in which the general base is a N-terminal group, displays a smaller
323 barrier than the standard mechanism in which His41 acts as the general base activating the water
324 molecule, as expected from the relative pK_a values (see Figure S4).
325
326 It is worth noticing that the peptide fragments obtained from this mechanistic proposal present a
327 salt-bridge between the charged C-terminal and N-terminal groups that must be broken during
328 products release. The separation of the two terminal groups implies that water molecules must
329 be placed, tightly bounded, between these two charged groups. Those configurations could
330 contribute to an inverse solvent isotope effect observed under steady state conditions only when

the scissile bond is an amide and not an ester (this is only when a N-terminal group is available to act as the general base).[24]

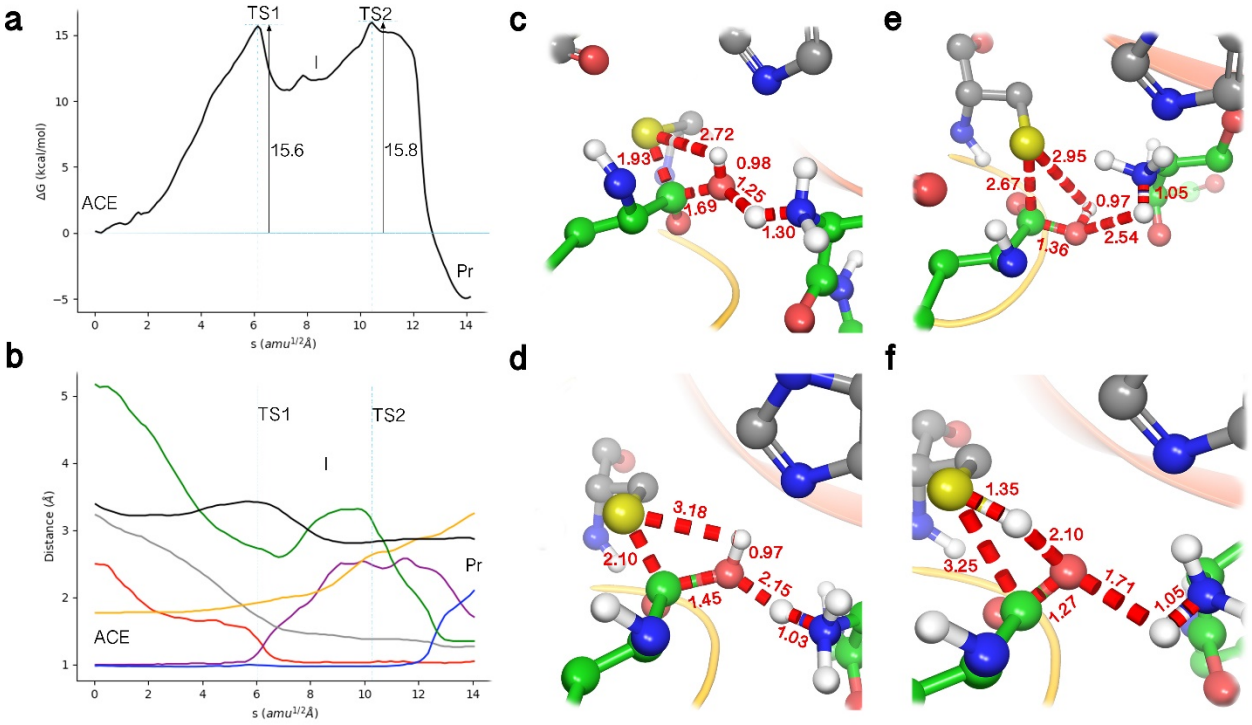


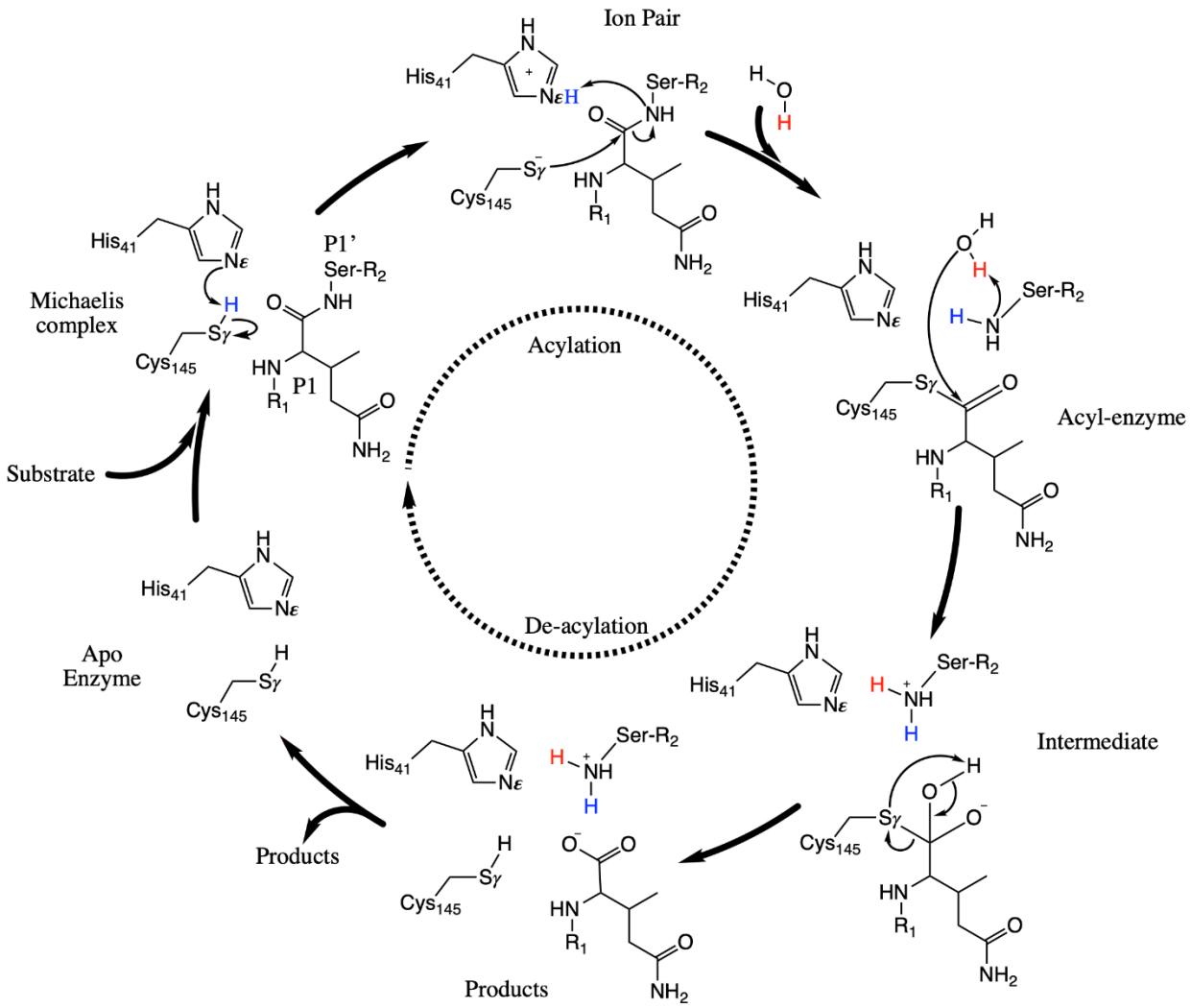
Figure 4. Simulation of the de-acylation reaction with the N-terminal group of the P' fragment acting as general base in charge of water activation. (**4a**) B3LYPD3/6-31+G*/MM free energy profile along the path-CV for the de-acylation reaction. The process takes place in two steps. In the first one the water activated by the N-terminal group attacks the acyl-enzyme complex (ACE) to form a thiodiolate intermediate (I) through the first TS (TS1). In the second one the reaction product (Pr) is obtained after breaking the acyl-enzyme bond in the second TS (TS2). (**4b**) Evolution of the distances selected as collective variables along the minimum free energy path. $S\gamma$-H_{w2} in green, $N\varepsilon$-N(P1') in black, O_w-C(P1) in grey, N(P1')-H_{w1} in red, $S\gamma$-C(P1) in yellow, O_w-H_{w1} in purple and O_w-H_{w2} in blue. (**4c**) Representation of TS1 where the water molecule is transferring a proton to the N-terminal group of the P' fragment and the resulting hydroxyl anion attacks the carbonyl carbon atom of the P fragment (C(P)). The values of the distances correspond to the coordinates of the MFEP at TS1. (**4d**) Representation of the thiodiolate intermediate (I). (**4e**) Representation of TS2 corresponding to the breaking of the $S\gamma$-C(P1) bond and the proton transfer from the carboxylic terminal group to the leaving sulfur atom. (**4f**) Representation of the reaction products (Pr) with the P-COO^- and P'-NH_3^+ peptide fragments in the active site of the protease.

Conclusions

We have presented a detailed analysis of the Michaelis complex and the proteolysis mechanism in the 3CL^{pro} of SARS-CoV-2 using DFT/MM computational simulations. Our study has identified key interactions established between the protein and a peptide substrate and the detailed reaction mechanism (see the complete reaction cycle in Figure 5). The reaction process involves the formation of a catalytic dyad ion pair from which the acylation step can be proceed (see Supplementary Video 1). In this acylation, the TS involves the proton transfer from His41 to the nitrogen atom followed by the nucleophilic displacement of the peptide bond by the Sγ atom of Cys145. For the de-acylation step (Supplementary Video 2), we have proposed a novel mechanism where the N-terminal group of the firstly formed peptide fragment is the general base catalyzing the hydrolysis of the acyl-enzyme complex. This mechanistic proposal can explain some of the experimental differences observed between amide and ester substrates in the highly similar protease of SARS-CoV. Our simulations stress on the role of the interactions established by the P1' moiety during the binding and reaction processes, indicating that this group can play an important role, both from thermodynamic and kinetic perspectives, in the design of better inhibitors of the 3CL^{pro} of SARS-CoV-2.

Acknowledgements

The authors acknowledge financial support from Feder funds and the Ministerio de Ciencia, Innovación y Universidades (project PGC2018-094852-B-C22). We also acknowledge PRACE for awarding us access to MareNostrum based in Spain at Barcelona Supercomputing Center (BSC). The support of Alejandro Soriano from Servei d'Informàtica de la Universitat de València, and Cristian Morales and David Vicente from BSC to the technical work is gratefully acknowledged. Authors also deeply acknowledge Dr. Kirill Zinovjev for assistance in the adaptation of the code and helpful discussions during about the use of the string method.



390
391 **Figure 5.** Schematic representation of the proteolysis mechanism in 3CL^{pro} of SARS-CoV-2 as deduced
392 from our simulations.
393
394

Methods

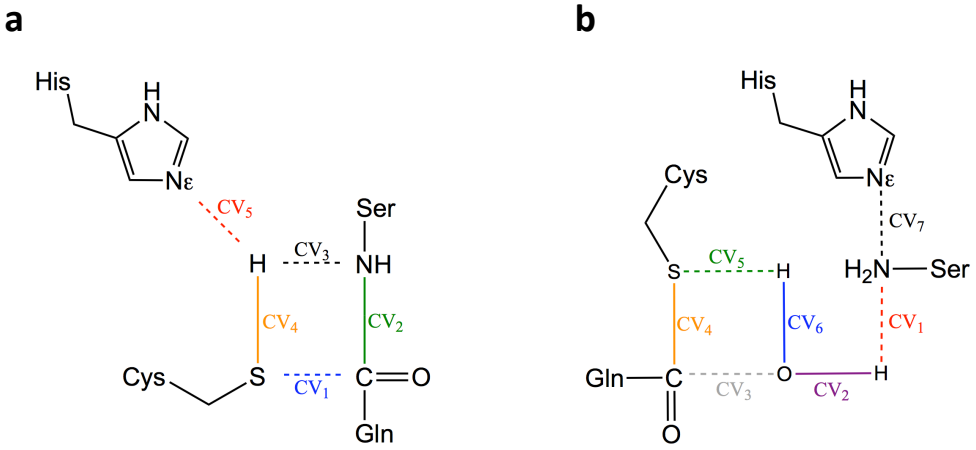
Classical Molecular Dynamics Simulations

The crystal structures with PDB codes 6Y2F[7] and 3AW0[30] were used as starting points to build the Michaelis complex. The former corresponds to the holo protease of SARS-CoV-2, the latter is the crystallographic structure of the SARS-CoV ortholog co-crystalized with the peptidic aldehyde inhibitor Ac-Ser-Ala-Val-Leu-His-Aldehyde. To build the Michaelis complex corresponding to the $3CL^{pro}$ protease of SARS-CoV-2 the two protein structures were aligned and then the co-crystalized ligand in 6Y2F was replaced with the crystallized ligand in 3AW0 in the two active sites of the homodimer (protomers A and B). The peptidomimetic Ac-Ser-Ala-Val-Leu-His-Aldehyde inhibitor was elongated using the Maestro tool[31] until we built the substrate like sequence Ac-Ser-Ala-Val-Leu-Gln-Ser-Gly-Phe-NMe in the two active sites. The absent hydrogens atom were added using the Protein Preparation Wizard tool of Maestro, and PROPKA3.0[32] was used to calculate the protonation states of titratable residues at pH 7.4.

The tleap tool from AmberTools18[33] was used to prepare the simulation systems. The Michaelis complexes, described with the ff14SB force field[34], were solvated into a box with a buffer region of at least 12 Å from any protein/substrate atom to the limits of the simulation box. TIP3P water molecules[35] were used. Na^+ atoms were added to neutralize the charge. The resulting system was minimized using 500 steps of steepest descent method followed by the conjugate gradient method, until the root mean square of the gradient was below 10^{-3} kcal $mol^{-1}Å^{-1}$. The system was then heated from 0 to 300 K using a heating rate of 1.7 $K \cdot ps^{-1}$. The backbone heavy atoms were restrained in a cartesian space using a harmonic potential with a force constant of 20 kcal $mol^{-1}Å^{-2}$. Along the equilibration step, the positional restraint force constant was changed from 15 to 3 kcal $mol^{-1}Å^{-2}$, decreasing by 3 units every 1.25 ns, and after 6.25 ns the positional restraints were removed and the systems continued their equilibration until 7.5 ns of NPT (300 K and 1 bar) simulation was completed. Then, 8 μs of NVT simulation at 300 K was performed with a 2 fs time step using SHAKE[36]. The Particle Mesh Ewald method was employed to describe the long range electrostatic interactions,[37,38] for the short range interactions a cutoff of 10 Å was used. Pressure was controlled by the Berendsen barostat and the temperature by the Langevin thermostat. For all the simulations, periodic boundary conditions were employed. The AMBER19 GPU version of pmemd[15,16] was used to run the classical molecular dynamic simulations. In order to sample a

427 reasonable configurational space of the protein, two additional replicas of the Michaelis complex
428 model with different initial velocities were run during 1 μs.
429
430 ***QM/MM Calculations***
431
432 Exploration of the free energy surfaces associated to the peptide bond breaking process have
433 been carried out using QM/MM simulations. In most simulations (see Table S1 for details), the
434 side chains of the catalytic dyad (Cys145 and His41) and a fragment of the peptide substrate
435 were included in the QM region while the rest of the system was described at the MM level as
436 explained above. The part of the substrate described at the QM level includes the two residues
437 involved in the peptide bond to be broken (Gln-P1 and Ser-P1') and the previous and next peptide
438 bonds up to the C^{α} atoms of Leu-P2 and Gly-P2'. In the exploration of the hydrolysis of the acyl-
439 enzyme we also included a water molecule in the QM region. To describe the QM subsystem we
440 used the B3LYP functional[39,40] and D3 dispersion corrections[41]. Calculations were performed with
441 the 6-31+G* basis set, unless indicated (we also employed the 6-31G* basis set). This level of
442 theory has been shown to be one of the best combinations to describe enzymatic reactions[42]. In
443 addition, this computational description for the QM region (B3LYPD3/6-31+G(d)) provides a gas
444 phase enthalpic change for the proton transfer reaction between imidazole and methanethiol (the
445 motifs of Cys and His side chains) in excellent agreement with the experimentally derived value
446 (see Supplementary Methods). All calculations were run with a modified version of Amber18[33,43]
447 coupled to Gausssian16[44] for Density Functional Theory calculations. A cutoff-radius of 15 Å was
448 used for all QM-MM interactions and the temperature was 300 K.
449
450 In order to explore the free energy landscape associated to the chemical reaction we used our
451 implementation of the string method, the Adaptive String Method (ASM).[18] In this method N
452 replicas of the system (the nodes of the string) are evolved according to the averaged forces and
453 kept equidistant, converging in such a way to the minimum free energy path (MFEP) in a space
454 of arbitrary dimensionality defined by the collective variables (CVs). Once the string has
455 converged, (see an example in Figure S5) we define a single path-CV (a collective variable called
456 *s*) that measures the advance of the system along the MFEP. This path-CV is used as the reaction
457 coordinate to trace the free energy profile associated to the chemical transformations under
458 analysis. The MFEPs were explored on a free energy hypersurface defined by a set of CVs formed

459 by those distances showing relevant changes during the process under study: H-Sγ(Cys145), H-
460 Nε(His41), H-N(P1'), Sγ(Cys145)-C(P1) and C(P1)-N(P1') for the acylation step (see Scheme 2a)
461 and H_{w1}-O_w, H_{w1}-N(P1'), O_w-C(P1), Sγ(Cys145)-C(P1), H_{w2}-O_w, N(P1')-Nε(His41) and H_{w2}-
462 Sγ(Cys145) for the de-acylation step (see Scheme 2b). Different initial guesses, corresponding to
463 different mechanistic proposals, were explored at the B3LYPD3/MM level using first the 6-31G*
464 basis set and then the best candidates were recalculated using the 6-31+G* basis set. Each of
465 the strings was composed of at least 96 nodes (see Supplementary Table S1), which were
466 propagated with a time step of 1 fs until the RMSD of the string felt below 0.1 $amu^{1/2}$·Å for at least
467 2 ps. The converged MFEPs are then averaged to define the *s* path-CV corresponding to each
468 string. The free energy profiles along the path CVs were obtained using an Umbrella Sampling
469 algorithm[45], running simulations for at least 10 ps and were integrated using WHAM technique[46].
470 The values of the force constants employed to bias the ASM simulations are determined on-the-
471 fly to ensure a probability density distribution of the reaction coordinate as homogeneous as
472 possible.[18] Replica exchange between neighbor string nodes was attempted every 50 steps to
473 improve convergence.
474
475 For the proton transfer between Cys145 and His41, considering the proximity of the proton donor
476 and acceptor atoms and the geometrical simplicity of the process, we traced the free energy
477 profile using Umbrella Sampling[45] along a simple proton transfer coordinate defined as the
478 antisymmetric combination of the distances of the proton to the donor and the acceptor atoms
479 (d(Nε-H)-d(Sγ-H)). For this profile only the side chains of the two involved residues were included
480 in the QM region (using the B3LYPD3/6-31+G* level of theory). A total of 40 windows were used,
481 corresponding to an increment of the reaction coordinate of 0.06 Å, each of them composed of
482 10 ps of equilibration and 20 ps of data collection. The force constant employed to drive the
483 reaction coordinate change was of 600 $kcal·mol^{-1}·Å^{-2}$. All the rest of details of the simulations were
484 as described before. We also used Umbrella Sampling along distinguished coordinates explore
485 the free energy landscape associated to the separation between the first peptide fragment and
486 the acyl-enzyme complex and between the two peptide fragments at the end of the de-acylation
487 process. Details of all the free energy simulations performed in this work are given in
488 Supplementary Table S1.
489
490
491


a
His
H
N
Nε
CV5
Ser
H
CV3
NH
CV4
CV2
Cys
S
CV1
C
O
Gln
b
His
H
N
Nε
Cys
CV7
S
CV5
H
H2N
Ser
CV4
CV6
CV1
Gln
C
CV3
O
CV2
H
O

Scheme 2

506 **References**
507

508 1. Zhou, P. *et al.* A pneumonia outbreak associated with a new coronavirus of probable bat
509 origin. *Nature* **125,** 270–273 (2019).
510 2. Lau, S. K. P. *et al.* Severe acute respiratory syndrome coronavirus-like virus in Chinese
511 horseshoe bats. *Proc. Natl. Acad. Sci.* **102,** 14040–14045 (2005).
512 3. Cheng, V. C. C., Lau, S. K. P., Woo, P. C. Y. & Yuen, K. Y. Severe Acute Respiratory
513 Syndrome Coronavirus as an Agent of Emerging and Reemerging Infection. *Clin.*
514 *Microbiol. Rev.* **20,** 660–694 (2007).
515 4. Pillaiyar, T., Manickam, M., Namasivayam, V., Hayashi, Y. & Jung, S.-H. An Overview of
516 Severe Acute Respiratory Syndrome–Coronavirus (SARS-CoV) 3CL Protease Inhibitors:
517 Peptidomimetics and Small Molecule Chemotherapy. *J. Med. Chem.* **59,** 6595–6628
518 (2016).
519 5. Bangham, C. R. M. The immune control and cell-to-cell spread of human T-lymphotropic
520 virus type 1. *J. Gen. Virol.* **84,** 3177–3189 (2003).
521 6. Jin, Z. *et al.* Structure of Mpro from COVID-19 virus and discovery of its inhibitors.
522 *bioRxiv* (2020). doi:10.1101/2020.02.26.964882
523 7. Zhang, L. *et al.* Crystal structure of SARS-CoV-2 main protease provides a basis for
524 design of improved α-ketoamide inhibitors. *Science (80-.).* **368,** 409 (2020).
525 8. Dai, W. *et al.* Structure-based design of antiviral drug candidates targeting the SARS-
526 CoV-2 main protease. *Science (80-.).* **368,** 1331 LP-1335 (2020).
527 9. Chen, Y. W., Yiu, C.-P. B. & Wong, K.-Y. Prediction of the SARS-CoV-2 (2019-nCoV) 3C-
528 like protease (3CLpro) structure: virtual screening reveals velpatasvir, ledipasvir, and
529 other drug repurposing candidates. *F1000Research* **9,** 129 (2020).
530 10. Rut, W. *et al.* Substrate specificity profiling of SARS-CoV-2 main protease enables design
531 of activity-based probes for patient-sample imaging. *bioRxiv* 2020.03.07.981928 (2020).
532 doi:10.1101/2020.03.07.981928
533 11. Hilgenfeld, R. From SARS to MERS: crystallographic studies on coronaviral proteases
534 enable antiviral drug design. *FEBS J.* **281,** 4085–4096 (2014).
535 12. Brocklehurst, K., Willenbrock, F. & Sauh, E. Cysteine proteinases. in *Hydrolytic Enzymes*
536 39–158 (Elsevier, 1987).
537 13. Jin, Z. *et al.* Structure of Mpro from SARS-CoV-2 and discovery of its inhibitors. *Nature*
538 **582,** 289–293 (2020).
539 14. Jin, Z. *et al.* Structural basis for the inhibition of SARS-CoV-2 main protease by
540 antineoplastic drug carmofur. *Nat. Struct. Mol. Biol.* **27,** 529–532 (2020).
541 15. Le Grand, S., Götz, A. W. & Walker, R. C. SPFP: Speed without compromise - A mixed
542 precision model for GPU accelerated molecular dynamics simulations. *Comput. Phys.*
543 *Commun.* **184,** 374–380 (2013).
544 16. Salomon-Ferrer, R., Götz, A. W., Poole, D., Le Grand, S. & Walker, R. C. Routine
545 microsecond molecular dynamics simulations with AMBER on GPUs. 2. Explicit solvent
546 particle mesh ewald. *J. Chem. Theory Comput.* **9,** 3878–3888 (2013).
547 17. Vanden-Eijnden, E. & Venturoli, M. Revisiting the finite temperature string method for the
548 calculation of reaction tubes and free energies. *J. Chem. Phys.* **130,** 194103 (2009).
549 18. Zinovjev, K. & Tuñón, I. Adaptive Finite Temperature String Method in Collective
550 Variables. *J. Phys. Chem. A* **121,** 9764–9772 (2017).
551 19. Fan, K. *et al.* Biosynthesis, purification, and substrate specificity of severe acute
552 respiratory syndrome coronavirus 3C-like proteinase. *J. Biol. Chem.* **279,** 1637–1642
553 (2004).
554 20. Yang, H. *et al.* The crystal structures of severe acute respiratory syndrome virus main

protease and its complex with an inhibitor. *Proc. Natl. Acad. Sci.* **100,** 13190 (2003).
21. Chen, H. *et al.* Only One Protomer Is Active in the Dimer of SARS 3C-like Proteinase. *J. Biol. Chem.* **281,** 13894–13898 (2006).
22. Barrila, J., Bacha, U. & Freire, E. Long-Range Cooperative Interactions Modulate Dimerization in SARS 3CLpro. *Biochemistry* **45,** 14908–14916 (2006).
23. Anand, K. *et al.* Structure of coronavirus main proteinase reveals combination of a chymotrypsin fold with an extra α-helical domain. *EMBO J.* **21,** 3213–3224 (2002).
24. Solowiej, J. *et al.* Steady-State and Pre-Steady-State Kinetic Evaluation of Severe Acute Respiratory Syndrome Coronavirus (SARS-CoV) 3CLpro Cysteine Protease: Development of an Ion-Pair Model for Catalysis. *Biochemistry* **47,** 2617–2630 (2008).
25. Świderek, K. & Moliner, V. Revealing the Molecular Mechanisms of Proteolysis of SARS-CoV-2 Mpro from QM / MM Computational Methods . Revealing the Molecular Mechanisms of Proteolysis of SARS-CoV-2 M pro from QM / MM Computational Methods . (2020). doi:10.26434/chemrxiv.12283967.v1
26. Hsu, W.-C. *et al.* Critical Assessment of Important Regions in the Subunit Association and Catalytic Action of the Severe Acute Respiratory Syndrome Coronavirus Main Protease. *J. Biol. Chem.* **280,** 22741–22748 (2005).
27. Elsässer, B. *et al.* Distinct Roles of Catalytic Cysteine and Histidine in the Protease and Ligase Mechanisms of Human Legumain As Revealed by DFT-Based QM/MM Simulations. *ACS Catal* **7,** 5585–5593 (2017).
28. Szawelski, R. J. & Wharton, C. W. Kinetic solvent isotope effects on the deacylation of specific acyl-papains. Proton inventory studies on the papain-catalysed hydrolyses of specific ester substrates: analysis of possible transition state structures. *Biochem. J.* **199,** 681–692 (1981).
29. Grimsley, G. R., Scholtz, J. M. & Pace, C. N. A summary of the measured pK values of the ionizable groups in folded proteins. *Protein Sci* **18,** 247–251 (2009).
30. Akaji, K. *et al.* Structure-based design, synthesis, and evaluation of peptide-mimetic SARS 3CL protease inhibitors. *J. Med. Chem.* **54,** 7962–7973 (2011).
31. Schrödinger Release 2016-3: Maestro, Schrödinger, LLC, New York, NY. (2016).
32. Olsson, M. H. M., Søndergaard, C. R., Rostkowski, M. & Jensen, J. H. PROPKA3: Consistent Treatment of Internal and Surface Residues in Empirical pKa Predictions. *J. Chem. Theory Comput.* **7,** 525–537 (2011).
33. Case, D. A. *et al.* Amber 2017. *Univ. California, San Fr.* AMBER 2017, University of California, San Francisc (2017). doi:citeulike-article-id:2734527
34. Maier, J. A. *et al.* ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB. *J. Chem. Theory Comput.* **11,** 3696–3713 (2015).
35. Jorgensen, W. L., Chandrasekhar, J., Madura, J. D., Impey, R. W. & Klein, M. L. Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* **79,** 926 (1983).
36. Ryckaert, J.-P., Ciccotti, G. & Berendsen, H. J. C. Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. *J. Comput. Phys.* **23,** 327–341 (1977).
37. Darden, T., York, D. & Pedersen, L. Particle mesh Ewald: An N·log(N) method for Ewald sums in large systems. *J. Chem. Phys.* **98,** 10089–10092 (1993).
38. Essmann, U. *et al.* A smooth particle mesh Ewald method. *J. Chem. Phys.* **103,** 8577–8593 (1995).
39. Becke, A. D. Density-functional thermochemistry. III. The role of exact exchange. *J. Chem. Phys.* **98,** 5648–5652 (1993).
40. Lee, C., Yang, W. & Parr, R. G. Development of the Colle-Salvetti correlation-energy

formula into a functional of the electron density. *Phys. Rev. B* **37,** 785–789 (1988).

41. Grimme, S., Antony, J., Ehrlich, S. & Krieg, H. A consistent and accurate ab initio parametrization of density functional dispersion correction (DFT-D) for the 94 elements H-Pu. *J. Chem. Phys.* **132,** 154104 (2010).
42. Shimato, T., Kasahara, K., Higo, J. & Takahashi, T. Effects of number of parallel runs and frequency of bias-strength replacement in generalized ensemble molecular dynamics simulations. *PeerJ Phys. Chem.* **1,** e4 (2019).
43. Zinovjev, K. String-Amber. *GitHub repository* (2019). Available at: https://github.com/kzinovjev/string-amber. (Accessed: 24th June 2020)
44. Frisch, M. J. *et al.* Gaussian 16 Rev. C.01. (2016).
45. Torrie, G. M. & Valleau, J. P. Nonphysical sampling distributions in Monte Carlo free-energy estimation: Umbrella sampling. *J. Comput. Phys.* **23,** 187–199 (1977).
46. Kumar, S., Rosenberg, J. M., Bouzida, D., Swendsen, R. H. & Kollman, P. A. THE weighted histogram analysis method for free-energy calculations on biomolecules. I. The method. *J. Comput. Chem.* **13,** 1011–1021 (1992).