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

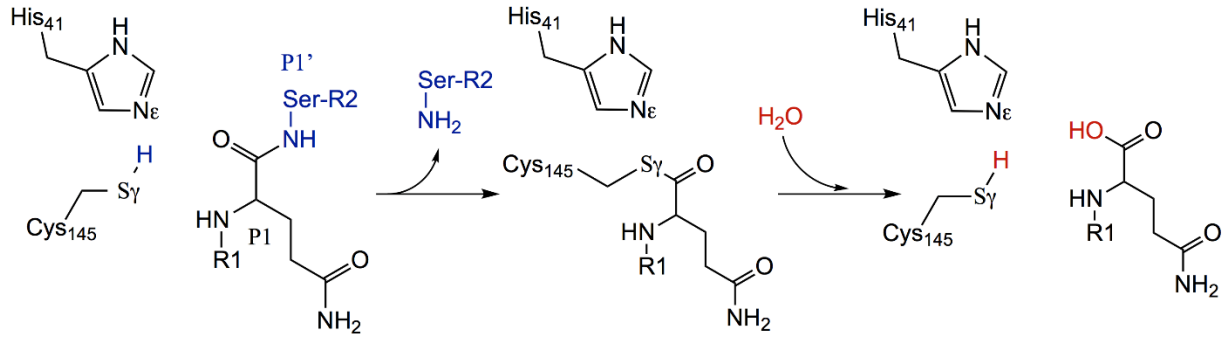
We present a detailed theoretical analysis of the reaction mechanism of proteolysis catalyzed by the main protease of SARS-CoV-2. Using multiscale simulation methods, we have characterized the interactions stablished by a peptidic substrate in the active site and then we have explored the free energy 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.

Introduction

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

The $3CL^{pro}$ enzyme of SARS-CoV-2 exists as a functional homodimer with two active sites in charge of cleaving the translated polyproteins into individual fragments to be used by the coronavirus[5]. As other cysteine proteases, each of the active sites contains a Cys-His catalytic dyad in charge of the hydrolysis of the peptide bond at specific sites of a polypeptide chain. Several structures of this protease have been already resolved by means of x-ray crystallography and deposited in the Protein Data Bank (PDB), including the free protease (PDB codes 6Y2E[2] and 6Y84[6]) and inhibitor bound proteases (PDB codes 6LU7[6] and 6Y2F[7]). The SARS-CoV-2 main protease has a structure virtually identical to the ortholog from SARS-CoV (96% identity). Even more, the main residues involved in catalysis, binding and dimerization processes are fully conserved[8]. Consequently, these two ortholog enzymes display highly similar substrate preferences[9]. The substrate cleavage by the $3CL^{pro}$ takes place between Gln at the P1 position and a Gly/Ala/Ser at the P1' one (see Scheme 1), being the presence of Gln an essential requirement[10].

In principle, the reaction mechanism of cysteine proteases involves two basic steps (see Scheme 1)[11]. 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. In fact, crystalized inhibitor bound structures resembles the acyl-enzyme complex[12,13].



Scheme 1

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. We have performed Molecular Dynamics (MD) simulations and the resulting structural models were then used to explore the reaction mechanism in this protease. Our work has allowed to unravel the molecular details of the reaction mechanism and to identify some key interactions established between the viral protease and the substrate at the Michaelis complex and at the reaction TS, in particular the role of the P1' residue. Identification of these interactions could be used to guide the design of future inhibitors of this enzyme.

Results

As detailed in the Methods section we carried out Molecular Dynamics 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 9 μs of classical simulations were run using the AMBER19 GPU version of *pmemd*[14,15]. 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[16,17] was employed to find the minimum free energy path (MFEP) on a multidimensional free energy surface defined by the distances of all the bonds to be broken or formed during the process and to trace the associated free energy profiles, without assuming *a priori* any particular mechanism.

Enzyme-Substrate Complex

The time evolution of the root mean square deviation (RMSD) values for each replica (2 protomers and 2 substrates in each replica) are shown in Figure S1. These values show that the protein structure is well-equilibrated and there are no large differences with respect to the initial structures, prepared from the x-ray data. Regarding the substrate, the presented RMSD values correspond small fluctuations in the active site. The observations made on the three replicas (one of 7 μs and two of 1 μs) are very similar in all cases. The substrate-binding pocket is divided into a series of subsites (denoted as S_i and S_i') each accommodating a single amino acid residue of the substrate placed before (P_i) and after (P_i') the peptide bond to be broken (the bond placed between Gln(P1- and Ser-P1' residues of the substrate). The map of hydrogen bond interactions observed during our MD simulations of the Michaelis complex is given in Figure 1a, a general view of the substrate in the two active sites of the dimer formed by protomers A and B is shown in Figure 1b, while an insight into the active site of protomer A is provided in Figure 1c. The $3CL^{pro}$ of SARS-CoV-2 (as is also the case of the SARS-CoV ortholog) presents a high specificity for Gln at P1 position[18]. As seen in Figure 1a the P1 residue is the one establishing more hydrogen bond interactions with the enzyme. The O and N main chain atoms of Gln-P1 are found to make hydrogen bonds with main chain atoms of Gly143, Ser144 and His164 (unless indicated all the residues of the protein belong to the same protomer A). Regarding the side chain of Gln-P1, this

is accommodated into the S1 subsite through hydrogen bond contacts with the main chain atoms of Phe140 and Leu141, with the Nε atom of His163 and the Oε atoms of Glu166. This last residue is in turn hydrogen bonded to the terminal NH group of Ser1 from protomer B. In fact, the N-terminal fragment (N-finger) of protomer B plays an active role pre-organizing the active site of protomer A for catalysis. This fact is in is in agreement with the observations made on the 3CL protease of related coronavirus[19], where it was found that dimerization is an essential condition for catalysis[18,20,21] and, consequently, those mutants lacking the N-finger fragment are almost completely inactive[22]. The side chain of Leu-P2 is surrounded by the side chains of His41, Met49, His164, Met165 and Asp187, while the main chain amide group is hydrogen bonded to the Oε atom of Gln189. The side chain of Val-P3 is solvent-exposed, while main chain NH and O atoms are hydrogen bonded to main chain atoms of Glu166.

The binding subsite for Ala-P4 is constituted by main chain interactions with Gln189 and Thr190, while the side chain of this residue is placed between the side chains of Met165, Leu167 and Pro168. The S5 subsite is formed by the side chain of Pro168 and by main chain atoms of Thr190. This description of the 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 an inhibitor bound in the active site[12,13]. Interestingly, 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 with those inhibitors in which the fragment corresponding to residues P1'-P3' either is released during the formation of the acyl-enzyme complex[13] or is shorter than in the case of our substrate peptide[12].

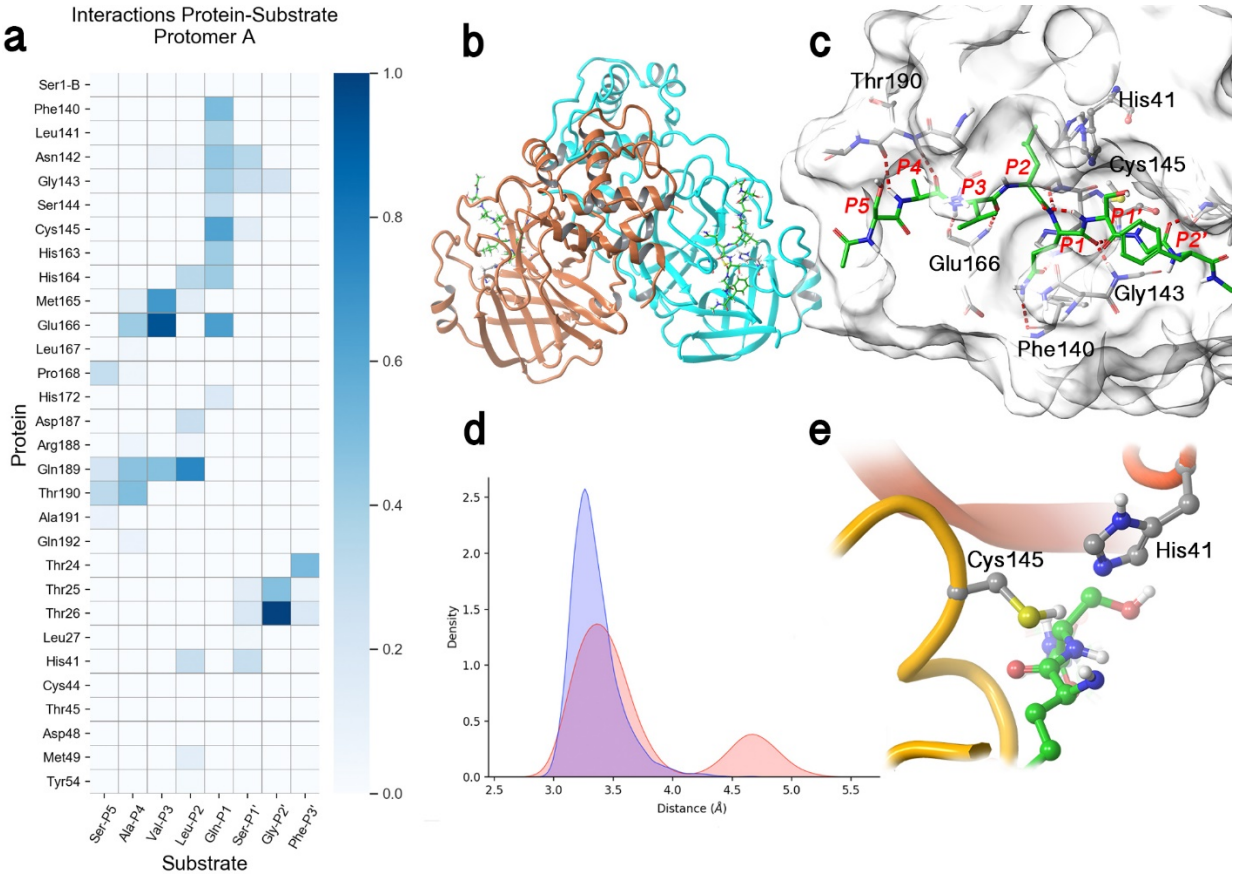


Figure 1. Results of the Molecular Dynamics simulation of 3CL protease of SARS-CoV-2 in complex with the substrate with sequence Ac-Ser-Ala-Val-Leu-Gln-Ser-Gly-Phe-NMe in the two active sites. (**1a**) Fraction of hydrogen bond contacts found during the trajectory of the Michaelis Complex between the residues of the substrate and those of the protease. 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 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.

The Catalytic Dyad

The reaction mechanism of cysteine proteases involves the nucleophilic attack of the Sγ atom of a cysteine (Cys145 in our case) to the C(P1) atom of the peptide bond. Figure 1d shows the probability distribution of Sγ-C(P1) distances found during our MD simulations. The distribution is clearly bimodal, with two peaks centered at 3.4 and 4.7 Å. These two peaks correspond to two different conformations of the side chain of Cys145, which can be present in *trans* and *gauche* conformations. This is in agreement with the observations made on the x-ray structure of the orthologue protease of SARS-CoV[19]. The most probable conformation corresponds to the *trans* conformer in which the Sγ sulfur atom is closer to the substrate (see Figure 1e). In both conformations the catalytic dyad remains hydrogen bonded, being the most probable distance between Cys145-Sγ and His41-Nε of about 3.3 Å (see Figure 1d). Interestingly, His41 is, in turn, hydrogen bonded, through a highly conserved crystallographic water molecule, to Asp187. This interaction can raise the pK_a of the histidine, increasing its ability to work as a base and abstract the proton from Cys145.

Considering the short distance observed between Cys145 and His41 and the possible activation of this last residue by the nearby Asp187, we explored the possibility to find the catalytic dyad forming an ion pair ($Cys^-/HisH^+$ or IP in Figure 2) instead of the neutral form modelled in the Michaelis complex (CysH/His or MC in Figure 2). We evaluated the free energy difference between these two forms of the dyad by means of free energy profiles associated to the proton transfer coordinate from Cys145 to His41 obtained at the B3LYPD3/MM level (see Methods). The proton transfer free energy profiles (Figure 2a) were calculated for the holo and apo forms of $3CL^{pro}$. According to Figure 2a the catalytic dyad is more stable in the neutral form, both for the apo and holo conformations. The catalytic dyad ion pair is 2.9 and 4.8 $kcal \cdot mol^{-1}$ above the neutral form, in the apo and holo enzymes, respectively. The protonated His41 is stabilized by hydrogen bond interactions with His164 and with Asp187, this last mediated by a crystallographic water molecule (see Figure 2b). The anionic Cys145 can be stabilized by the presence of water molecules or by the hydroxyl group of Ser-P1' in the holo form and by water molecules in the apo form. Bulkier residues at the P1' position could hinder the access of water molecules and thus stabilize the unprotonated form of the Cys145. This could be one of the factors explaining the preference of $3CL^{pro}$ for small residues at the P1' position (Ser/Ala/Gly). According to our free

energy profiles shown in Figure 2a the ion pair form of the catalytic dyad is better stabilized with respect to the neutral one in the apo enzyme than in the holo one (by almost 2 kcal·mol^{-1}), which can be related to the better solvation of the anionic cysteine in the former. In 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 ion pair 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[23].

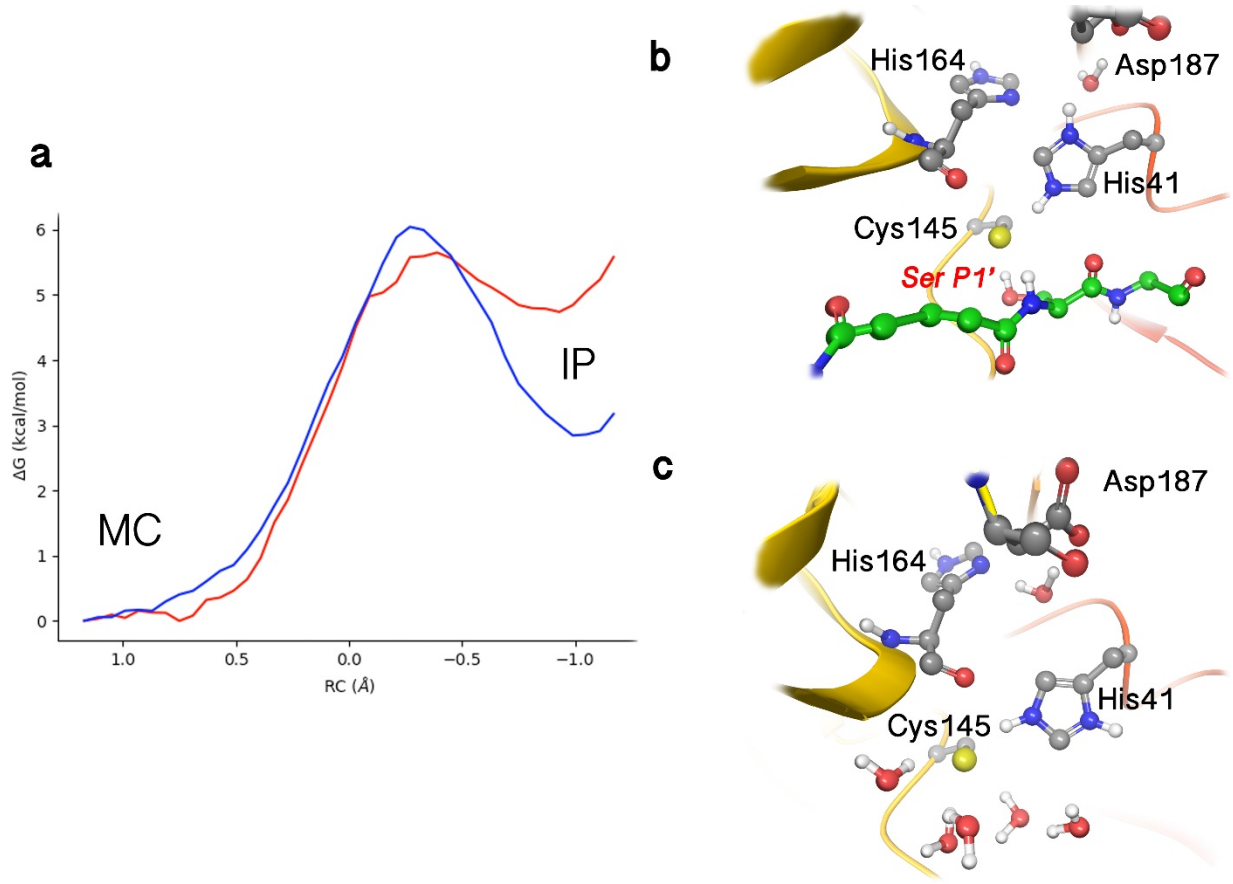


Figure 2. Analysis of the formation of an ion pair (IP) catalytic dyad (Cys145^{-}···His41H^{+}) from the neutral form (Cys145H···His41) found in the Michaelis Complex (MC). (**2a**) B3LYPD3/6-31+G* Free energy profile associated to the proton transfer coordinate from the Sγ atom of Cys145 to the Nε 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, with indication of 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.

The Acylation step

We explored the free energy landscape for the formation of the acyl-enzyme starting from the catalytic dyad ion pair at the B3LYPD3/MM level with the 6-31+G* basis set (and 6-31G* when indicated). The converged MFEP is shown in Figures 3a-b. According to our simulations, after ion pair formation, the acylation proceeds by means of a proton transfer from His41 to the N(P1') atom followed by the nucleophilic attack of Cys145-Sγ on the C(P1) atom and the simultaneous breaking of the C(P1)-N(P1') peptide bond. These elementary events take place in a concerted but asynchronous way. The TS found for this mechanism (see Figure 3c) is associated to the proton transfer from His41 to the amide nitrogen atom of the peptide bond N(P1'). At the TS the Sγ atom of Cys145 approaches to the C(P1) atom, reducing the interatomic distance from 3.11 to 2.34 Å. This approach is accompanied by a moderate lengthening of the peptide bond (the C(P1)-N(P1') distance being lengthened from 1.36 to 1.54 Å). According to the free energy path shown in Figure 3a, the total free energy barrier associated to the acylation process, including the free energy cost of the ion pair formation, is of 14.6 kcal·mol^{-1}. This value is compatible with the activation free energies derived from the steady-state rate constants measured at 25º C for peptidic substrates that are cleaved at the Gln-Ser bond by the highly similar ortholog 3CL^{pro} of SARS-CoV (between 16.2 and 17.2 kcal·mol^{-1})[23]. It must be noticed that in this proteolysis the acylation step is not considered to be the rate-limiting one and then these experimental activation free energies provide an upper limit for the acylation barrier[23]. Figure 3c shows that this TS is stabilized by means of a hydrogen bond interaction with the hydroxyl group of SerP1', indicating, also in agreement with the experimental observations, that the leaving group plays an important role in catalysis. Remarkably, the proposed reaction mechanism is also in good agreement with the experimental proton inventory results, that indicate that there are two protons in flight during the acylation, one at the transition state and another one at earlier stages[23]. In our picture these two proton transfers events correspond to the proton transferred from His41 to the N(P1') atom of the substrate at the TS and the proton transfer from Cys145 to His41 during the ion pair formation. Regarding the acyl-enzyme product, our free energy profile shows two possible conformations that differ in less than 1 kcal·mol^{-1}. The first conformer in the path displays a C(P1)-N(P1') distance of 2.4 Å, while in the second one the distance is increased up to 3.1 Å due to an inversion of the amino group that is now able to receive a hydrogen bond from a water molecule (see Figure 3d and S2). As discussed below, this water molecule will play a key role in the de-

acylation step. Finally, in our free energy profile the formation of the acyl-enzyme (see Figure 3d) is almost thermo-neutral, with a reaction free energy of about -1 kcal·mol^{-1}.

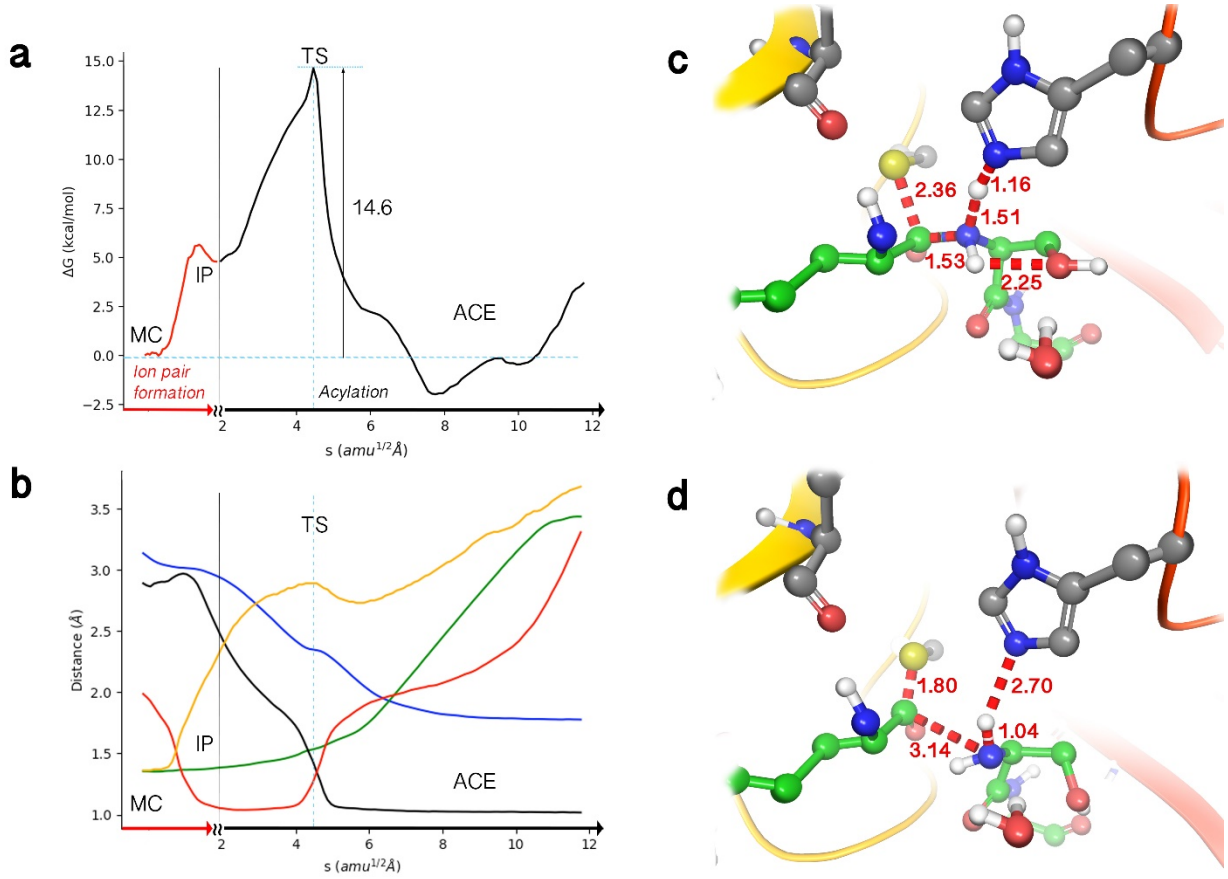


Figure 3. Simulation of the acylation reaction taking place through the formation of the ion pair. (**3a**) B3LYPD3/6-31+G* free energy profile along the path-CV for the acylation reaction after the formation of the (IP) from the Michaelis complex (MC). The reaction takes place with a single transition 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, S 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 discussed in the Supplementary Information (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[24]. 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[9]. Interestingly, this rate constant (0.050 s^{-1})[9] 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})[23]. 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[23,25]. 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 proposed in the literature for other cysteine proteases[26] 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). In this mechanism a neutral Cys145 directly attacks the peptide bond, transferring the proton to the N(P1') atom, forming a Sγ-C(P1) bond and breaking the C(P1)-N(P1') bond. This mechanism takes place in a single asynchronous step, where the proton transfer precedes the attack of the sulfur atom to C((P1). However, the free energy barrier associated to this reaction mechanism was found to be significantly higher than the experimental estimation.

The De-acylation step

Regarding the de-acylation step, the standard mechanistic proposal suggested for the related enzymes[27] 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 in a neutral form) and 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 S2). This configuration suggests an alternative reaction mechanism that can yield the two peptide fragments with correct protonation states in the terminal groups and regenerate the enzymatic active site in its most stable state (the neutral catalytic dyad). An additional advantage of this novel proposed mechanism is that involves water activation by means of the N-terminus of the P' fragment, which is known to be a better base than histidine side chains (the average pK_a values are about 7.7 and 6.6, respectively)[28]. Finally, as discussed below, the proposed mechanism can explain some of the experimental observations that only appear when the scissile bond is amide and not when the substrate is an ester (where a basic N-terminus is not formed after the acylation).[23] We obtained the MFEP corresponding to such a mechanism at the B3LYPD3/MM level (see Figure 4). The process involves three chemical stages: i) the proton transfer from the water molecule to the terminal amino group of the leaving P' peptide, resulting in the formation of the P'-NH_3^+ peptide fragment and the nucleophilic attack of the resulting hydroxide anion to the carbonyl carbon atom C(P1) atom; ii) the breaking of the C(P1)-Sγ bond to yield the P-COOH peptide, and iii) the proton transfer from the P-COOH group to Cys145, regenerating the enzyme in the correct protonation state and leaving the C-terminal group of the P peptide fragment unprotonated, allowing to complete the full catalytic cycle of the enzyme. This process is stepwise, presenting two TSs (see Figure 4a). The first TS (TS1 in Figure 4c) corresponds to the proton transfer from the water molecule to the N(P1') atom and the nucleophilic attack of the hydroxyl anion to the carbonyl carbon atom. The free energy barrier associated to this step is of 15.6 kcal·mol^{-1}, very close to the values derived from the reaction rate constants for the hydrolysis of peptides containing the Gln-Ser bond by the ortholog protease of SARS-CoV (from 16.2 to 17.2 kcal·mol^{-1}).[23] This proton transfer is concomitant to the attack of the hydroxyl group on the C(P1) carbonyl carbon atom,

resulting in the formation of an intermediate thiodiolate (Figure 4d). After rotation of the hydroxyl group to orient the proton towards the sulfur atom, the reaction proceeds with the cleavage of the C(P1)-Sγ bond. The second TS observed during the diacylation (TS2, see Figure 4e), corresponds to the separation of the Sγ atom (the C(P1)-Sγ distance being 2.67 Å). The free energy barrier associated to this second step from the acyl-enzyme complex is very close to the first one, 15.7 kcal·mol^{-1}. Afterwards, the leaving cysteine is stabilized by means of a proton transfer from the P-COOH group to the Sγ atom, regenerating the enzyme in its more stable protonation state (a neutral catalytic dyad) and yielding the P peptide fragment with a terminal carboxylate (the product is represented in Figure 4e). The proposed mechanism, in which the general base is a N-terminal group, displays a smaller barrier than the standard mechanism in which His41 acts as the general base activating the water molecule, as expected from the relative pK_a values (see Figure S4).

The products obtained from this mechanistic proposal present a salt-bridge between the two peptide fragments. Product release then requires the separation of the C- terminal and N-terminal groups, which are present in the active site in its charged states. The classical free energy profile traced for the separation of the two terminal groups (see Figure S5) shows that, at the TS, water molecules must be placed, tightly bounded, between these two charged groups. Remarkably, this TS configuration could be then the origin of inverse solvent isotope effects, observed under steady state conditions only when the scissile bond is an amide and not when the substrate is an ester, this is only when charged products can be formed through the proposed de-acylation mechanism.[23] Furthermore, the additional step due to the separation of the two resulting peptides could explain the kinetic complexity observed during the de-acylation step in the ortholog protease of SARS-CoV, only when the substrate is an amide and not when it is an ester.[23]

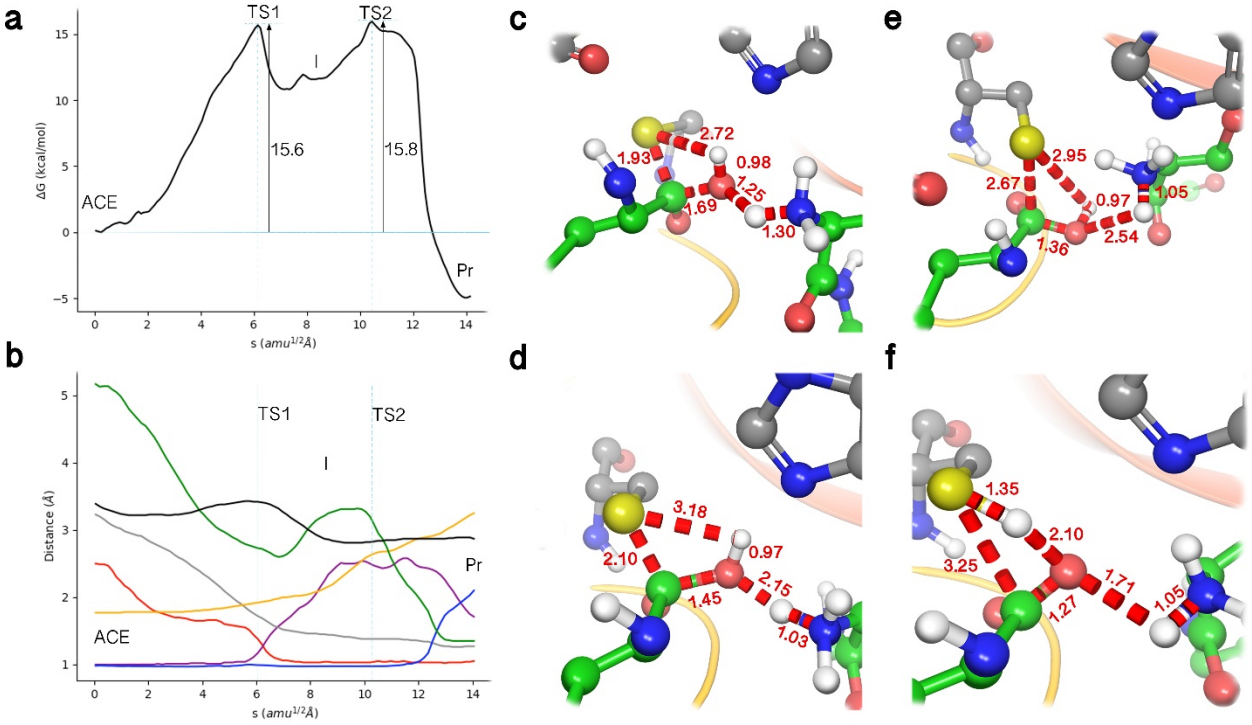


Figure 4. Simulation of the deacylation reaction with the N-terminus group of the P' fragment acting as general base in charge of water activation. (**4a**) B3LYPD3/6-31+G* free energy profile along the path-CV for the deacylation reaction. The process takes place in two steps. In the first one the water activated by the N-terminal group of the P' fragment 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). (**3b**) Evolution of the distances selected as collective variables along the minimum free energy path. Sγ-H_{w2} in green, H-Nε in red, Nε-N(P1') in black, O_w-C(P1) in grey, N(P1')-H_{w1} in red, Sγ-C(P1) in yellow, O_w-H_{w1} in purple and O_w-H_{w2} in blue. (**3c**) Representation of TS1 where the water molecule is transferring a proton to the N-terminus 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. (**3d**) Representation of the thidolate intermediate (I). (**3e**) Representation of TS2 corresponding to the breaking of the Sγ-C(P1) bond and the proton transfer from the carboxylic acid to the leaving sulfur atom. (3f) 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 proteolysis in the 3CL protease of SARS-CoV-2. This study shows that, in agreement with the kinetic characterization of the ortholog of SRAS-CoV, the process involves the formation of a catalytic dyad ion pair from which the acylation step can be proceed. In this acylation, the TS involves the proton transfer from His41 to the nitrogen atom of the scissile amide bond, while the subsequent nucleophilic displacement of the peptide bond by the attack of the Sγ atom of Cys145 is barrierless. For the deacylation step we have proposed a novel mechanism where the N-terminal group of the first peptidic fragment is the general base catalyzing the hydrolysis of the acyl-enzyme complex. This N-terminal group is more basic than the side chain of His41, that has been traditionally proposed as the activating base. This new mechanistic proposal can explain some of the experimental differences observed between amide and ester substrates in the hydrolysis catalyzed by highly similar protease of SARS-CoV. Our simulations stress on the role of the interactions established by the P1' moiety, both during the binding and reaction process, indicating that leaving groups play an important role, both from thermodynamic and kinetic perspective, in the design of better inhibitors of the action of the 3CL protease 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.

Methods

Classical Molecular Dynamics Simulations

The crystal structures with PDB codes 6Y2F[7] and 3AW0[29] 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[30] 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[31] was used to calculate the protonation states of titratable residues at pH 7.4.

The tleap tool from AmberTools18[32] was used to prepare the simulation systems. The Michaelis complexes, described with the ff14SB force field[33], 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[34] 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 - $Å^2$. Along the equilibration step, the positional restraint force constant was changed from 15 to 3 kcal/mol - $Å^2$, decreasing by 3 units every 1.25 ns, 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, 7 μs of NVT simulation at 300 K was performed with a 2 fs time step using SHAKE[35]. The Particle Mesh Ewald method was employed to describe the long range electrostatic interactions,[36,37] for the short range interactions a cutoff of 10 $Å^2$ 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[14,15] was used to run the classical molecular dynamic simulations. In order to sample a

reasonable configurational space of the protein, two additional replicas of the Michaelis complex model with different initial velocities were run during 1 μs.

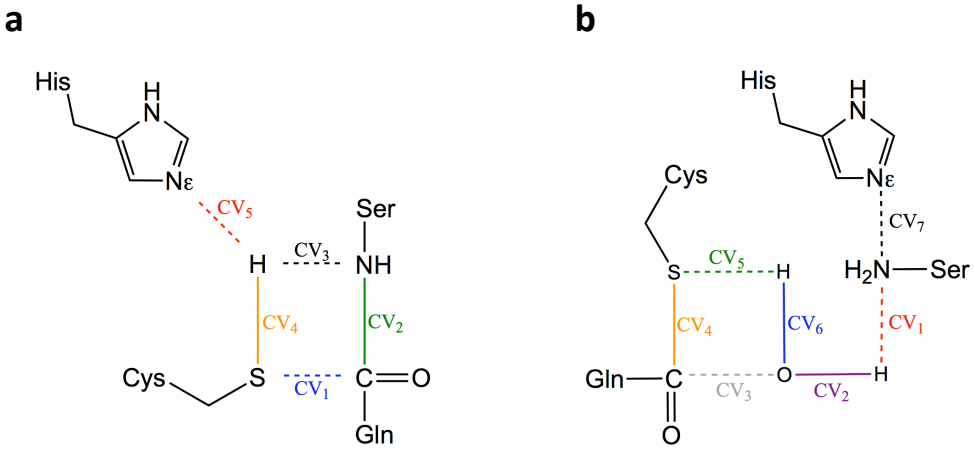
QM/MM Calculations

Exploration of the free energy surfaces associated to the peptide bond breaking process have been carried out using QM/MM simulations. In most simulations (see Table S1 for details), the side chains of the catalytic dyad (Cys145 and His41) and a fragment of the peptide substrate were included in the QM region while the rest of the system was described at the MM level. The part of the substrate described at the QM level includes the two residues involved in the peptide bond to be broken (Gln-P1 and Ser-P1') and the previous and next peptide bonds up to the C^{α} atoms of Leu-P2 and Gly-P2'. In the exploration of the hydrolysis of the acyl-enzyme we also included a water molecule in the QM region. To describe the QM subsystem we used the B3LYP functional[38,39] and D3 dispersion corrections[40]. Calculations were performed with the 6-31+G* basis set, unless indicated. This level of theory has been shown to be one of the best combinations to describe enzymatic reactions[41]. In addition, this computational description for the QM region (B3LYPD3/6-31+G(d)) provides a gas phase enthalpic change for the proton transfer reaction between imidazole and methanethiol (the motifs of Cys and His side chains) in excellent agreement with the experimentally derived value (see Supplementary Material, section S1). All calculations were run with a modified version of Amber18[32,42] coupled to Gausssian16[43] for DFT calculations. A cutoff-radius of 15 Å was used for all QM-MM interactions.

In order to explore the free energy landscape associated to the chemical reaction we use our implementation of the string method, the Adaptive String Method (ASM).[17] In this method N replicas of the system (the nodes of the string) are evolved according to the averaged forces and kept equidistant, converging to the minimum free energy path (MFEP) in a space of arbitrary dimensionality defined by the number of collective variables (CVs). Once the string has converged, we define a single path-CV (a collective variable called *s*) that measures the advance of the system along the MFEP. This path-CV is used as the reaction coordinate to trace the free energy profile associated to the chemical transformations under analysis. The MFEPs were explored on a free energy hypersurface defined by a set of CVs formed by those distances showing relevant changes during the process under study: H-Sγ(Cys145), H-Nε(His41), H-N(P1'),

Sγ(Cys145)-C(P1) and C(P1)-N(P1') for the acylation step (see Scheme 2a) 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}-Sγ(Cys145) for the de-acylation step (see Scheme 2b). Different initial guesses, corresponding to different mechanistic proposals, were explored at the B3LYPD3/MM level using first the 6-31G* basis set and then best candidates were recalculated using the 6-31+G* basis set. Each of the strings was composed of at least 96 nodes (see Table S1), which were propagated with a time step of 1 fs until the RMSD of the string felt below 0.1 $amu^{1/2}$·Å (see Figure S7). The MFEPs obtained are then averaged to define the *s* path-CV corresponding to each string. The free energy profiles along the path CVs were obtained using an Umbrella Sampling algorithm[44], running simulations for at least 10 ps and were integrated using WHAM technique[45]. The values of the force constants employed to bias the ASM simulations are determined on-the-fly to ensure a probability density distribution of the reaction coordinate as homogeneous as possible.[17]

For the proton transfer between Cys145 and His41, considering the proximity of the proton donor and acceptor atoms and the geometrical simplicity of the process, we traced the free energy profile using Umbrella Sampling[44] along a simple proton transfer coordinate defined as the antisymmetric combination of the distances of the proton to the donor and the acceptor atoms (d(Nε-H)-d(Sγ-H)). For this profile only the side chains of the two involved residues were described at the QM level (B3LYPD3/6-31+G*). A total of 40 windows were used, corresponding to an increment of the reaction coordinate of 0.06 Å, each of them composed of 10 ps of equilibration and 20 ps of data collection. The force constant employed to drive the reaction coordinate change was of 600 $kcal·mol^{-1}·Å^{-2}$. All the rest of details of the simulations were as described before. We also used Umbrella Sampling along distinguished coordinates explore the free energy landscape associated to the separation between the first peptide fragment and the acyl-enzyme complex and between the two peptide fragments at the end of the de-acylation process. Details of all the free energy simulations are given in Table S1.


a
His
H
N
Nε
CV5
CV3
Ser
H
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
O
CV3
O
CV2
H

Scheme 2

References

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

Biol. Chem. **281,** 13894–13898 (2006).
21. Barrila, J., Bacha, U. & Freire, E. Long-Range Cooperative Interactions Modulate Dimerization in SARS 3CLpro. *Biochemistry* **45,** 14908–14916 (2006).
22. 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).
23. 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).
24. 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
25. 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).
26. 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).
27. 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).
28. 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).
29. Akaji, K. *et al.* Structure-based design, synthesis, and evaluation of peptide-mimetic SARS 3CL protease inhibitors. *J. Med. Chem.* **54,** 7962–7973 (2011).
30. Schrödinger Release 2016-3: Maestro, Schrödinger, LLC, New York, NY. (2016).
31. 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).
32. Case, D. A. *et al.* Amber 2017. *Univ. California, San Fr.* AMBER 2017, University of California, San Francisc (2017). doi:citeulike-article-id:2734527
33. 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).
34. 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).
35. 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).
36. 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).
37. Essmann, U. *et al.* A smooth particle mesh Ewald method. *J. Chem. Phys.* **103,** 8577–8593 (1995).
38. Becke, A. D. Density-functional thermochemistry. III. The role of exact exchange. *J. Chem. Phys.* **98,** 5648–5652 (1993).
39. 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).
40. 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).
41. 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).
42. Zinovjev, K. String-Amber. *GitHub repository* (2019).
43. Frisch, M. J. *et al.* Gaussian 16 Rev. C.01. (2016).
44. Torrie, G. M. & Valleau, J. P. Nonphysical sampling distributions in Monte Carlo free-energy estimation: Umbrella sampling. *J. Comput. Phys.* **23,** 187–199 (1977).
45. 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).