UNIVERSITAT
JAUME I

Ph.D. Thesis

# Computational Study on the Reactivity and Inhibition of Arginine Gingipain B, a Potential Target for the Treatment of Alzheimer's Disease

*Author:*
Santiago Movilla Núñez

*Supervisors:*
Vicent Moliner Ibáñez
Mª Teresa Roca Moliner

Castelló de la Plana, Spain

**January** 2023

Doctoral Program in Theoretical Chemistry and Computational Modelling

Univesitat Jaume I – Doctoral School

# Computational Study on the Reactivity and Inhibition of Arginine Gingipain B, a Potential Target for the Treatment of Alzheimer's Disease

Report submited by Santiago Movilla Núñez in order to be eligible for a doctoral degree awarded by the Universitat Jaume I

Santiago Movilla Núñez

Firmado por
SANTIAGO
MOVILLA NUÑEZ -
NIF:Y7493571R
el día
09/01/2023 con

(Ph.D. Candidate)

Dr. Vicent Moliner Ibáñez

VICENTE
MOLINER
IBAÑEZ -
NIF:25388050Y

Firmado digitalmente
por VICENTE MOLINER
IBAÑEZ -
NIF:25388050Y
Fecha: 2023.01.08
15:34:22 +01'00'

(Supervisor)

Dr. María Teresa Roca Moliner

MARIA TERESA|
ROCA|MOLINER

Firmado
digitalmente por
MARIA TERESA|
ROCA|MOLINER
Fecha: 2023.01.04
11:16:54 +01'00'

(Supervisor)

Castelló de la Plana (Spain), January 2023

**FUNDING:**

- Generalitat Valenciana through a Santiago Grisolía Ph.D. grant GRISOLIAP/2019/064.
- HPC-Europa3 project through a grant of mobility HPC17RZ6NO.

## LIST OF PUBLICATIONS:

(1) Movilla, S.; Martí, S.; Roca, M.; Moliner, V. Unrevealing the Proteolytic Activity of RgpB Gingipain from Computational Simulations. *J. Chem. Inf. Model.* **2021**, 61 (9), 4582–4593. https://doi.org/10.1021/acs.jcim.1c00666.

(2) Movilla, S.; Martí, S.; Roca, M.; Moliner, V. Computational Study of the Inhibition of RgpB Gingipain, a Promising Target for the Treatment of Alzheimer's Disease. *J. Chem. Inf. Model.* **2023**. *Accepted.*

(3) Movilla, S.; Roca, M.; Moliner V.; Magistrato, A. Molecular Basis of RNA-Driven ATP Hydrolysis in DExH-box Helicases. *Submitted.*

*To the people who have supported*
*me in the process*

# ABSTRACT

Since its identification in 1901 by the German psychiatrist Alois Alzheimer, Alzheimer's disease has emerged as one of the greatest challenges in pharmaceutical research. Being the most common type of dementia, it presents severe effects such as loss of memory, cognitive abilities and physical dexterity, symptoms that occur mainly in the elderly population. Classified as a multifactorial disease, the attempt to develop a treatment for Alzheimer's disease has transited over a large number of molecular targets. In 2019, the cysteine protease RgpB was identified as a possible new pharmacological target for the development of neuroprotective treatments using small drugs-like molecules. In fact, a family of irreversible inhibitors was discovered and patented, and they are nowadays in advanced stages of clinical trials. All of the above highlights the need for a thorough and detailed understanding of possible targets for the design of drugs for Alzheimer's disease treatments.

This thesis represents an effort to shed light on atomic-level details of the mechanisms by which the cysteine protease RgpB works and the processes of its inhibition by small-molecules drugs candidates. For this purpose, a comprehensive computational study combining methods based on classical molecular dynamics and hybrid quantum mechanics/molecular mechanics (QM/MM) molecular dynamics was carried out. These methods were employed in different methodological schemes to obtain reaction free energies and binding free energies. This, allowing us to elucidate the mechanism by which the proteolysis reaction proceeds on RgpB, to characterize non-covalent interaction profiles between putative inhibitors and RgpB and to unravel the covalent binding mechanisms of such irreversible inhibitors.

In more detail, a preliminary step using classical molecular dynamics and principal component analysis (PCA) determined the most likely protonation state of the Cys/His catalytic dyad of RgpB and its influence in promoting the reactive arrangement of catalytic residues and peptide. Next, umbrella sampling QM/MM calculations were employed to evaluate all possible pathways of RgpB-catalyzed proteolysis. The most favorable mechanism was found to proceed through three steps. For acylation, initially, the sulfur atom of the Cys244 residue attacks the carbonyl carbon of the peptide and the proton of the Cys244 residue is transferred to the amino group of the peptide in a concerted manner. Then, the peptide bond is cleaved and a fragment is released. Finally, in the deacylation step a water molecule attacks the carbonyl carbon of the peptide and a proton from the water is

transferred to the Cys244 residue. For the mechanism found, the rate-limiting step free energy barrier is in very good agreement with the available experimental evidence. Notably, it should be emphasized that our hypothesized mechanism shows an unusual role of the His211 residue and a crucial role of the peptide in activating catalysis.

For the study of the inhibition of RgpB, a set of previously reported irreversible inhibitors was used as a reference. From these, by means of alchemical transformations and Poisson-Boltzmann surface area molecular mechanics (MM/PBSA) calculations supplemented with interaction entropies, the binding energies were calculated. The interactions governing the affinity of the inhibitors to the binding pocket were characterized by means of contact maps and interaction energies decomposed by residue. Notably, residues Ser213, Glu214, Asp158 and Asp281 were found to crucially interact via hydrogen bonds with the inhibitors. Finally, following the same methodology as for the wild-type substrate, the reaction mechanism by which these drug candidates lead to covalently bound complexes inhibiting the RgpB protease was evaluated. This process is performed in a single step, that consists of the activation/deprotonation of Cys244 by the oxygen atom of the carbonyl group concertedly with the attack of the S$\gamma$:Cys244 atom to the carbon atom of the same group.

In a final stage of the thesis, and in order to explore methodologies to enrich our findings on the activity and inhibition of RgpB, the use of metadynamics in enzyme reactivity studies was performed. This incursion was carried out on the investigation of the ATPase activity of Prp2 and its RNA regulation. We found that hydrolysis proceeds by a four-step mechanism in which the rate-determining step is the nucleophilic attack on the gamma phosphate. Complementary, MD simulations disclosed the molecular terms of RNA-driven activation of ATP hydrolysis. These results demonstrated the potential and applicability of DFT-based metadynamics in the study of biological systems, and in particular on systems that lower-level Hamiltonians can represent a source of error during the conformational sampling.

All in all, the work presented in this thesis contributes to the understanding of the reaction mechanism and inhibition of an enzyme that is a potential target for the treatment of Alzheimer's disease. This understanding will allow to rationalize and optimize future inhibitors in the pharmaceutical race for the development of new potential treatments.

# CONTENTS

# CHAPTER 1. INTRODUCTION

## 1.1 ENZYMATIC CATALYSIS

Enzymes are amazing catalysts that make vital processes viable in the time required for life. Mostly built from the same synthetic building blocks, called amino acids, they participate directly or indirectly in all biological processes.[1] For 2020, SwissProt/Uniprot[2,3] reported 20365 screened human proteins of which 3428 were fully characterized enzymes.[4]

Enzymes manage to increase the rate of some reactions up to $10^{23}$-fold compared to reactions in solution, this only with cellular enzymatic concentrations between micro- and millimolar.[5] Additionally, they often have switch *on/off* activity allowing them to regulate their catalytic capacity according to cellular requirements.[5] Since its discovery, enzymatic catalysis has been in the focus of interest in several areas of chemistry as they carry out chemical reactions with an efficiency, versatility, enantioselectivity and specificity hardly matched by synthetic catalysts.

On the other hand, given their extensive participation in metabolic processes, their function or dysfunction is associated with the development of multiple diseases. Likewise, through controlled regulation by inhibitors (drugs), treatments and cures for these diseases can be developed. However, the full chemical and pharmaceutical potential requires a thorough understanding of the details that make enzyme catalysis or inhibition possible.[6]

**Explaining enzyme catalytic efficiency.** The catalytic efficiency of enzymes has been of great interest to chemistry. In energetic terms it can always be concluded that enzymes drive reactions through lower activation barriers ($\Delta G_U$ vs $\Delta G_C$) (Figure 1.1).[7] This is either because within the enzyme the reaction goes through another reaction mechanism or because directly its barrier is lower than its equivalent in solution.[7]

Understanding how they are able to influence system energetics and accelerate the reaction rate by several orders of magnitude is crucial to exploit their applications.

However, given the immense variety of enzymes and the enormous variety of reactions they catalyze, a single explanation is insufficient. The main proposals that explain enzyme catalysis are shown in the following list.[5]

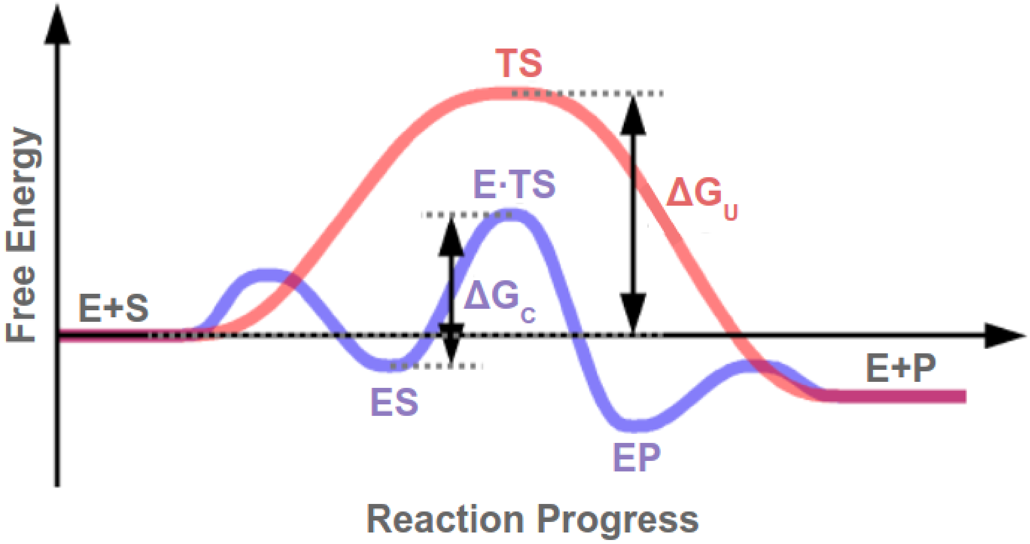


***Figure 1.1.*** *Schematic representation of the free energy profile for an enzyme-catalyzed reaction (purple) and the corresponding reaction in solution (pink). $\Delta G_U$ represents the uncatalyzed reaction free energy barrier and $\Delta G_c$ represent the enzyme-catalyzed free energy barrier. ES and EP represents the enzyme:substrate and enzyme:products complexes, respectively.*

a) *Stabilization of the transition state:*[8–14] this idea states that the enzyme has the ability to interact with the transition state in such a way that it is stabilized. Basically, this is achieved by the enzyme being complementary to the transition state in terms of geometrical and electrostatic effects. In this case, in order to result in a net lowering of the activation barrier, the enzyme:substrate complex is less stabilized than the transition state.

b) *Intrinsic binding energy:*[15,16] this effect is particularly influential on multi-substrate reactions. The effects of binding energy come from the sum of all non-covalent interactions between enzyme and substrates, including all van der Waals interactions associated with structural complementarity as well as desolvation and electrostatic. All these interactions make the binding enthalpy high enough to balance the unfavorable entropy of bringing two or more molecules together. Thus, compensating for the entropic term that in solution is not redressed at the time of reactive collision.

c) *Reacting group approximation, orientation and orbital steering:*[17–19] stereo-electronic assistance basically boils down to i) converting multi-substrate reactions into "unimolecular" reactions ii) increasing the local effective concentration of the reactants with respect to each other and iii) effectively arranging and orienting the reactive functional groups not only in terms of proximity but also in electronic/orbital orientation, analogously to the orbital symmetry conservation rules.[20,21] This, from a kinetic perspective, increases the number of effective collisions by several orders of magnitude.

d) *Reactant state destabilization:*[15,22] in this case, it is proposed that the enzyme is able to induce strain or tension in certain substrates by arranging them in conformations that they would not explore in solution. This destabilization increases the energy of the reactants and therefore reduces the activation barrier necessary for catalysis. The most widely accepted explanation for this phenomenon is that the interactions between the enzyme and the non-reactive parts of the substrate are so favorable that they compensate for the enzymatic environment forcing destabilization of the reactive center of the substrate. This is broadly known as the *Circe Effect.*[15] Other branches of this hypothesis such as *Near-Attack Conformers*[23,24] are also recognized.[12]

e) *Covalent catalysis:*[25–27] as the name implies, in this case the kinetics of a reaction is increased by the covalent participation of the enzyme in the reaction. This basically promotes the formation of transient, more reactive intermediates, discretizing the reaction into better defined stages. These stages allow the mechanism to be organized in a way that entropic changes are managed while conserving and taking advantage of the transfer potential of the reactive groups. These mechanisms usually involve the functional groups of lysines, histidines, cysteines, serines or threonines in the active sites of the enzymes.

f) *Acid/base catalysis:*[5] the formation of formal cations or anions is an unlikely event that requires very specific conditions to take place. Processes that result in the localization of formal charges on individual organic atoms usually have very high activation barriers. Acids and bases enhance reactivity preventing the formation of these charges by donating or withdrawing protons, respectively. Enzymes enhance the reaction rate by making acidic or basic groups readily available for catalysis. In solution, when available, acids or bases must coincide at the exact moment when the other points of the reaction are also reactively disposed. The constant availability of acidic or basic residues of the enzyme increases the probability of such reactive events. These groups can be histidines, lysines, glutamates or aspartates present in the active sites. When their protonation state is not restored in the same reaction, they can do so later with the solvent.

g) *Metal ions facilitation:*[5] although the effect of a metal results in a mixture of several of the above factors, it is easier to categorize them independently when they are a consequence of the presence of a metal ion. Taking advantage of the well-defined coordination spheres of metals, some enzymes are able to orient substrates around them in highly reactive ways. In addition, the electrostatic environment and their Lewis acid nature modify the reactivity of the reactive centers by modifying their electronegativities or their pKas. Thus, metals have particular properties that allow to increase the reaction rate.

h) *Promoting catalysis by enzyme dynamics:*[28–33] In this factor, the enzyme acts more integrally as the large macromolecule that it is. By means of protein dynamics such as vibrations or conformational changes, they carry out processes that would otherwise be unfeasible or extremely unlikely. Enzymes acting on large molecules such as other proteins, RNA or DNA can behave as molecular machines/motors that coordinate intrincate physical or chemical changes. Likewise, even in small molecule reactions it is considered that protein dynamics contribute to force catalysis. Primarily, the whole enzyme vibrates in the direction of the reaction change, inducing the faster formation of the transition state.

### 1.1.1 ENZYME KINETICS

Enzymes (E) are well-defined molecular structures. As explained before, the mechanisms they use to increase reaction rates are well defined and in all cases depend on close contact with their substrate (S). The result of forming such specific interactions, with the substrate correctly oriented, is what is known as an enzyme:substrate (ES) complex. In a first base model, developed by L. Michaelis and M. L. Menten,[34,35] in which all intermediates and transition states are omitted, a single substrate enzyme catalysis reaction can be represented as:

$$\mathrm{E+S} \underset{k_{-1}}{\overset{k_1}{\rightleftharpoons}} \mathrm{ES} \underset{k_{-2}}{\overset{k_2}{\rightleftharpoons}} \mathrm{E+P} \quad \text{(Eq. 1.1)}$$

where the k's represent the kinetic constants for the direct and reverse reactions of the first and second steps. For such a chemical expression, one can define the rate law[36] in terms of products as:

$$v=\frac{d[P]}{dt}=k_2[ES]-k_{-2}[E][P] \quad \text{(Eq. 1.2)}$$

It can be thought that by measuring the kinetics of the reaction in early stages, the product concentration will be close to zero ($[P]\approx 0$) and therefore the above equation becomes:

$$v=\frac{d[P]}{dt}=k_2[ES]\ (\text{Eq. 1.3})$$

However, for this expression to be analytically useful, we must express it in terms of the concentration of reagents. Values for which we have control experimentally. For this purpose, we will introduce several assumptions that will simplify the model:

a) The formation of the enzyme:substrate complex is fast and therefore the equilibrium between E, S and ES species is reached quickly.

b) The catalytic step $ES \rightleftharpoons E+P$ is slow and is therefore the rate-determining step.

c) Given points I and II, as soon as ES is converted into products it is quickly reestablished by the first equilibrium $E+S \rightleftharpoons ES$. Thus keeping the concentration [ES] constant over the measurement time. This is known as steady-state approximation.

d) $[S]\gg[E_T]$ so the concentration [ES] at equilibrium is not significant compared to that [S]. So once equilibrium $E+S \rightleftharpoons ES$ is reached, at the beginning of the reaction, we can assume that the concentration [S] has not changed much ($[S]_{eq}\approx[S]_o$).

e) The enzyme exists only in two states, free and enzyme:substrate complex. Therefore the concentration $[E_T]$ is the sum of $[E_T]=[E]+[ES]$.

Under these assumptions it is much simpler to derive an analytical equation of the reaction rate in terms of reactants. For this purpose, we will use the steady-state aproximation to [ES], to introduce the relation

$$\frac{d[ES]}{dt}=k_1[E][S]-k_{-1}[ES]-k_2[ES]+k_{-2}[E][P]=0$$
$$k_1[E][S]+k_{-2}[E][P]=k_{-1}[ES]+k_2[ES] \text{ (Eq. 1.4)}$$

Again, if the measurement is made in the early stages of the reaction, the product concentration is close to zero ($[P]\approx 0$) and the equation becomes:

$$k_1[E][S]=k_{-1}[ES]+k_2[ES] \text{ (Eq. 1.5)}$$

If we group the constants in a single term we obtain the expression:

$$[E][S]=\left(\frac{k_{-1}+k_2}{k_1}\right)[ES] \text{ (Eq. 1.6)}$$

Where the ratio term between the constants is what is known as the Michaelis constant, $K_M$[34,35]

$$K_M=\frac{k_{-1}+k_2}{k_1} \text{ (Eq. 1.7)}$$

Using the assumption e) and introducing $K_M$ in the equation 1.6:

$$([E_T]-[ES])[S]=[ES]K_M \text{ (Eq. 1.8)}$$

and finally clearing the term [ES] from the equation, we have:

$$[E_T][S]-[ES][S]=[ES]K_M$$
$$[E_T][S]=[ES]K_M+[ES][S]$$
$$[E_T][S]=[ES](K_M+[S])$$

$$\frac{[E_T][S]}{K_M+[S]}=[ES] \text{ (Eq. 1.9)}$$

And now, we can simply substitute the expression of [ES] into the equation 1.3:

$$v_0=\frac{k_2[E_T][S]}{K_M+[S]} \text{ (Eq. 1.10)}$$

Where the subscript 0 in $v_0$ indicates that the expression is valid only at the beginning of the reaction. This equation is analytically useful since it is in terms of initial concentrations that are determinable.

In order to further restructure the equation we can imagine the situation in which the concentration of S, [S], is sufficiently high. This condition has both chemical and mathematical consequences. On the one hand, chemically, a high substrate concentration S shifts the equilibrium $E+S \rightleftharpoons ES$ towards ES. This implies that almost all of the enzyme is in the ES state ($E_T \approx ES$) and then [ES] is in its maximum value. Likewise, we can deduce from equation 1.3 that the velocity is also approaching the maximum achievable ($v \approx v_{max}$). This phenomenon is called enzyme saturation, and refers to the fact that all the available enzyme is catalyzing ( $ES \rightleftharpoons E+P$) and an extra addition of substrate would not increase the velocity.

On the other hand, mathematically, if [S] reaches very high values, such that $[S] \gg K_M$, then the $K_M+[S]$ term of the equation 1.10 can be approximated to [S]. Resulting in the expression

$$v_{max}=\frac{k_2[E_T][S]}{[S]}=k_2[E_T] \text{ (Eq. 1.11)}$$

Finally, substituting $v_{max}$ into the equation 1.10 gives the Michaelis-Menten equation:

$$v_0 = \frac{v_{max}[S]}{K_M + [S]} \text{ (Eq. 1.12)}$$

This can be rewritten in the form:

$$\frac{1}{v_0} = \frac{K_M}{v_{max}} \frac{1}{[S]} + \frac{1}{v_{max}} \text{ (Eq. 1.13)}$$

This is the best known and most practical form of the Michaelis-Menten equation. Its potential lies in the fact that simple experimental measurements of initial velocity as a function of substrate concentration allow the kinetic parameters $K_M$ and $v_{max}$ to be obtained. In turn, if we rewrite the equation 1.11 as:

$$v_{max} = k_2[E_T] = k_{cat}[E_T] \text{ (Eq. 1.14)}$$

From where, if you know $v_{max}$ and the initial enzyme concentration, $[E_T]$, $k_{cat}$, can be calculated. This parameter is called the turnover number.

**Significance of $K_M$ and $k_{cat}$, and Eyring equation.** $K_M$ in practical terms turns out to be a concentration. The substrate concentration, [S], at which the system reaches half the maximum velocity $v_{max}$. Thus, we can think that the lower its value, the lower the concentration needed to reach this velocity ($v_{max}/2$). Thus, at a low $K_M$, a small amount of substrate is sufficient to reach half saturation of the enzyme. In other words, the substrate is highly affine to the enzyme and has a very high tendency to form the enzyme:substrate, ES, complex. By this reasoning we, can deduce that $K_M$ is inversely proportional to the affinity of the substrate for the enzyme. The higher the $K_M$ value, the lower the affinity of the substrate for the enzyme and the higher the concentration of substrate required to saturate the enzyme.

The $k_{cat}$, meanwhile, indicates the number of substrate molecules that a single active site can convert into products in one second at a given enzyme concentration

$[E]_0$. As previously implied, this is easily obtained by equation 1.14, $k_{cat}=V_{max}/[E_T]$. Even more interesting is its equivalence to $k_2$. Since $k_2$ represents the kinetic constant for the conversion of the enzyme:substrate complex, ES, it is closely related to the efficiency of catalysis. In fact, $k_{cat}$ ($k_2$) being the kinetic constant of the chemical step, one can use the Transition State Theory (TST)[37] to relate it to the energetics of the process. Specifically, by means of the Eyring equation (equation 1.15), similar to the experimental version of Arrhenius, the value of the free energy of the activation barrier at a given temperature can be calculated from the value of the kinetic constant, $k_{cat}$.

$$k=\frac{\kappa k_B T}{h} e^{\frac{-\Delta G^{\ddagger}}{RT}} \quad \text{(Eq. 1.15)}$$

where κ represents the transmission coefficient (which is usually close to one except for processes that involve ligth atom transfers, such as proton or hydride transfers). In this way, the turnover number, $k_{cat}$, often allows computational chemists to validate the simulations performed with the experimental measurements. Since the activation free energy, and the chemical stage with which it is associated, is one of the parameters of greatest interest in the field.

### 1.1.2 ENZYME INHIBITION

Enzyme inhibition is the process by which a molecule other than the natural substrate (may be the product) impedes catalysis.[5] This can be classified in either reversible or irreversible inhibition. Reversible is when the free/unbound inhibitor and the bound inhibitor states are in chemical equilibrium. While in irreversible inhibition the reverse reaction after the binding of the inhibitor is practically null.[5] Within reversible inhibition we find competitive, non-competitive and uncompetitive inhibition.

**Competitive Inhibition.** In competitive inhibition a molecule binds to the active site of the enzyme, at the same site where the substrate binds. The molecule (inhibitor) is usually structurally and chemically similar to the substrate, thus capable of binding to the same active site. Because of the binding, the competitive inhibitor blocks the active site and competes with the substrate for the place, but its effects can be reduced by increasing the substrate concentration. Competitive inhibitors show an increase in $K_M$ due to the competition that results in an apparent decrease in affinity of the substrate for the enzyme. The $v_{max}$ is not affected since once the substrate enters the catalytic process is unaffected and proceeds at the same rate.

**Non-competitive inhibition.** Non-competitive inhibition involves the binding of a molecule to a site other than the active site (an allosteric site). The binding of the inhibitor to the allosteric site induces a conformational change in the active site of the enzyme resulting in the active site and substrate no longer sharing optimal complementarity, meaning that the substrate can bind but not react. Since the inhibitor does not compete directly with the substrate, increasing substrate levels cannot mitigate the effect of the inhibitor. In this case since the enzyme:substrate complex is not reactive but can form without problem as the active site is free, a decrease in $v_{max}$ is observed but $K_M$ remains the same.

**Uncompetitive inhibition.** The remaining, uncompetitive inhibition consists in the anchoring of the inhibitory molecule in the enzyme:substrate complex, ES, that blocks the catalytic capacity of the enzyme. The anchoring of this molecule can occur either in the vicinity of the active site or in allosteric pockets. Since the reactive ability of the ES complex is reduced, the $v_{max}$ is also reduced. The effect on $K_M$ is a bit more complex to explain but we can use Le Châtelier's principle to do so. Since the concentration of the ES complex decreases as it enters in a second equilibrium with the enzyme:substrate:inhibitor complex, ESI, the equilibrium with the free enzyme and substrate shifts to fill this vacancy. As a result, equilibrium $E+S \rightleftharpoons ES$ favors more the ES state appearing to have a higher substrate affinity and resulting in a decrease in $K_M$.

Irreversible inhibitors, in turn, can be categorized into those that bind covalently or non-covalently to the target enzyme. Those that bind non-covalently to the enzyme function similarly to the reversible ones but at equilibrium the dissociation of the inhibitor is slow due to highly specific interactions. Those that bind covalently to the enzyme, on the other hand, are a whole new world and will be described in more detail below as they are of particular interest to this work.

**Covalent inhibition.** Covalent inhibitors are molecules designed to form a covalent bond with a specific enzyme target.[38] Depending on the nature of the bond, which is usually controlled by the selection of the reactive warhead, one can have reversible or irreversible covalent inhibitors. Different types of warheads have been developed over the years, aiming to be selective initially to react with specific amino acids.[39] The amino acids targeted today in the design of covalent inhibitors are cysteine, serine, threonine, tyrosine and lysine. There are inhibitors targeting other residues but these are really rare.

The design of covalent inhibitors is the constant search for a balance between their potency, given by their reactivity, and the toxicity resulting from promiscuity with undesirable targets.[40] On one hand, covalent inhibitors present several advantages such as incomparable efficacy with non-covalent inhibitors and thus lower dose requirements, or a marked reduction in the possibility of drug resistance and targeting of shallow binding sites. All these advantages comes come from the potency of their interaction with the enzyme. For example, by covalently binding to the enzyme they are highly potent, with low $IC_{50}$ values and long binding duration. As a consequence, the required dose and frequency of intake for these drugs are lower compared to non-covalent drugs.[41] In turn, by reacting (usually) with residues that are essential for enzyme activity, covalent inhibitors have been reported to have potential advantages against resistance-prone targets, given the inability of the enzyme to mutate these residues.[42]

On the other hand, covalent inhibitors have disadvantages that must be carefully addressed during their design. Because they react with amino acid residues (present in all proteins) they may react with non-target enzymes.[43] These side reactions result in unexpected toxicity or hypersensitivity. In turn, being largely irreversible, they may not be suitable for targets that are rapidly degraded, degrading and excreting the inhibitors as well.[39] However, with a detailed knowledge of the action mechanism and the reactivity of the warheads, the design of covalent inhibitors can be properly addressed by means of refining their properties to minimize their disadvantages.

**Covalent inhibitors categorization:** As mentioned at the beginning, covalent inhibitors, like non-covalent inhibitors, can be reversible or irreversible. The reversible ones can be either reversible covalent type or slow substrate type.[40] The former, as their name indicates, form a covalent adduct that can be unbonded again. Described by the equation:

$$\mathrm{E+I} \overset{\mathrm{K_I}}{\rightleftharpoons} \mathrm{E\cdot I} \overset{\mathrm{k_{inac}}}{\rightleftharpoons} \mathrm{E-I} \text{ (Eq. 1.16)}$$

Where the interpunct denotes non-covalent interaction and the dash denotes covalent bond.

Slow substrate-type inhibitors are not properly reversible, but are denoted as such by intermittently releasing the enzyme (as in an equilibrium). Basically, these inhibitors form a covalent adduct with the enzyme which is then catalyzed into a product. This catalysis step is very slow, resulting in very long occupancy times of the inhibitor on the enzyme and consequent inhibition.[44]

$$\mathrm{E+I} \overset{\mathrm{K_I}}{\rightleftharpoons} \mathrm{E\cdot I} \overset{\mathrm{k_{inac}}}{\rightleftharpoons} \mathrm{E-I} \xrightarrow{\text{slow}} \mathrm{E+P} \text{ (Eq. 1.17)}$$

Moreover, irreversible inhibitors are divided into three categories:

a) *Residue-specific reagent:* The first class is more a tool in *in vitro* biochemical studies than potential drugs.[40] Residue-specific, is basically used to react with specific amino acids. But in addition to distinguishing between amino acids, they are unable to distinguish between targets.[45–48]

b) *Affinity label:*[49] similar to the first, is used for labeling but targeting the molecule to more specific sites, this is achieved by the addition of groups that, through non-covalent interactions, map to specific pockets. In some cases, these groups that direct the compound to ligand binding sites can subsequently be released leaving the fragment covalently bound on the enzyme.[50–52] In this category we find most of the covalent type drugs, basically, compounds are designed that through non-covalent interactions are affine to the target enzymes, forming an enzyme-substrate complex that subsequently reacts by binding covalently and irreversibly to the enzyme.[53]

$$\mathrm{E+I} \overset{\mathrm{K_I}}{\rightleftharpoons} \mathrm{E \cdot I} \overset{\mathrm{k_{inac}}}{\rightarrow} \mathrm{E-I} \quad \text{(Eq. 1.18)}$$

c) *Mechanism-based:* the last group, mechanism-based inhibitors,[54] binds to the active site of enzymes by the normal catalytic mechanism to produce reactive species that results in covalent bond formation.

$$\mathrm{E+I} \overset{\mathrm{K_I}}{\rightleftharpoons} \mathrm{E \cdot I} \overset{\mathrm{k_{cat^*}}}{\rightleftharpoons} \mathrm{E \cdot I^*} \overset{\mathrm{k_{inact}}}{\rightarrow} \mathrm{E-I^*} \quad \text{(Eq. 1.19)}$$

Where I* represents a derivative of I that was produced after standard catalysis and will react to covalently and irreversibly bind to the enzyme forming E-I*. This class of inhibitors are the most selective, as they do not depend only on non-covalent interactions to reach their target, but also do not become active until after reacting in the same way as the natural substrate. This double filter gives them a selectivity hardly achieved by other classes of covalent inhibitors.

## 1.2 PROTEOLYTIC ENZYMES

IUPAC defines proteases as all those enzymes that catalyze the hydrolysis of proteins or peptides.[36] Although they were initially conceived as mere destructive and aggressive proteins, progress in research has demonstrated their ability to hydrolyze highly specific substrates, producing new protein products essential for life. In fact, the current research apogee of these proteins is due to the large amount of investigations that have linked them to the control of multiple metabolic processes, and consequently their malfunction is associated with diseases. Their roles are present in DNA replication and transcription, cell proliferation and differentiation, tissue morphogenesis and remodeling, heat shock responses, angiogenesis, neurogenesis, hemostasis, blood coagulation, inflammation, immunity, autophagy, senescence, necrosis and apoptosis, among many other roles. Given this, proteases are often in the focus of pharmaceutical research as potential targets for drug and treatment design or as potential diagnostic biomarkers.[55–60]

The protease landscape is enormous. The curated Degradome Database,[61] which does not include pseudogenes or retrovirus-derived sequences, reports to date 588 proteases in humans and 672 in mice. This enormous number of proteases are classified according to the mechanism of catalysis. By this classification we find six groups of proteases.[62] On the one hand, glutamic, aspartic and metalloproteases, in which the nucleophilic attack on the peptide bond is carried out directly by a water molecule or a derived hydroxyl ion. On the other hand, there are serine, cysteine and threonine proteases, in which the nucleophile is the side chain of the respective amino acid. It should be noted that in mammals no glutamic proteases have been reported so far and therefore the interest in these is a little less. In humans, the list is headed by metalloproteases and serine proteases with 192 and 184 proteins reported.[61] This is followed by cysteine proteases with 164 reported and more distant threonine and aspartic proteases with 27 and 21 members identified, respectively.

### 1.2.1 CYSTEINE PROTEASES

As mentioned above, proteases where the nucleophile attacking the peptide bond is the sulfanyl group of a cysteine residue are called cysteine-type proteases or just cysteine proteases. They share several similarities with serine and threonine proteases, in both their active site structure and general reaction mechanism.[62] The main characteristics that all cysteine proteases share are:

a) *Acyl-enzyme formation:*[63–66] The adduct obtained as an intermediate after the attack of the sulfanyl group on the scissile peptide bond, and in which the protein is covalently bound to one of the fragments, is known as an acylenzyme. Its formation has been extensively verified both chemically and kinetically. Reactions catalyzed by cysteine proteases always proceed through the following general mechanism:

$$\mathrm{E+S} \rightleftharpoons \mathrm{ES} \rightleftharpoons \mathrm{EA}\ (-\mathrm{P1}) \rightleftharpoons \mathrm{E+P2} \quad \text{(Eq. 1.20)}$$

Where E, S, ES, EA, P1 and P2 correspond respectively to the enzyme, the substrate, the enzyme-substrate complex, the acylenzyme, the first released protein fragment and the second protein fragment. The stage of formation of EA and P1 is known as acylation, while the next stage is known as deacylation. While this general scheme is shared among cysteine proteases, they often diverge in the mechanistic details of the stages such as the general basis, protonation states, and the steps into which each stage is divided (i.e. stepwise or concerted).

b) *Imidazole group dependency:*[67–69] In all cysteine proteases reported so far, a histidine residue is present in the active site. Indeed, evidence from kinetic studies with mutants shows that the imidazole group of His is essential for catalysis. It is common to find the Cys/His dyad arranged in close spatial proximity, complementing their protonation states and acting as acid/base as the stage of the reaction requires. However, crystallographic structures and mechanistic studies show that they may be interacting via water molecules or even separated by the substrate.

c) *Oxyanion binding site:*[70–73] The oxyanionic hole corresponds to a region found around the substrate in cysteine, serine and threonine proteases. This region is composed of the NH groups of the backbone of two residues of the enzyme, which interact via hydrogen bonds with the oxygen atom of the carbonyl group of the peptide bond to be broken. The role of these groups is both to increase the electronegativity in the carbonyl of the ES state and to stabilize the charges located on the oxygen atom in the transition state and the acylenzyme. Although it is always present, kinetic studies show that unlike in the case of serine or threonine proteases, in cysteine proteases this structure is not essential and reactivity is not greatly affected when the interaction with the oxyanionic hole is disturbed.

d) *Presence of a third acid residue:*[74–78] In most of cysteine proteases a Cys/His catalytic dyad is sufficient for catalysis, although many of them show decreased activity upon mutation of a third residue (usually an Asp or a Glu). These residues, always negatively charged, are found interacting via hydrogen bonds with the imidazole group of the catalytic histidine. In some cases it modulates the acidity/basicity of the imidazole group, while in others its role is simply to orient it correctly. In less frequent cases, this group can be found close to cysteine, even participating directly in the reaction mechanism.

Despite their convergence into these features, cysteine proteases are classified into 12 clans, including the unassigned “clan”. The largest of these clans is the CA, to which the well-known and widely studied papin belongs.[62] The second largest known clan is the CD clan, where clostripain, legumain, caspase-1 and the gingipains are classified. As for enzymes in general, the classification by clan is done by means of fold patterns and distribution of the domains (quaternary structure).

#### 1.2.1.1 Clan CD and Gingipain Family

The CD clan is composed of cysteine proteases with a characteristic α,β,α-sandwich folding (Figure 1.2). The sequence motifs preceding the histidine and the catalytic cysteine correspond to two independent blocks of hydrophobic residues. The histidine is always followed by a glycine, of which its backbone NH participates in the oxyanionic hole. The sequence containing the histidine is located in a solvent-exposed loop.[79–83]

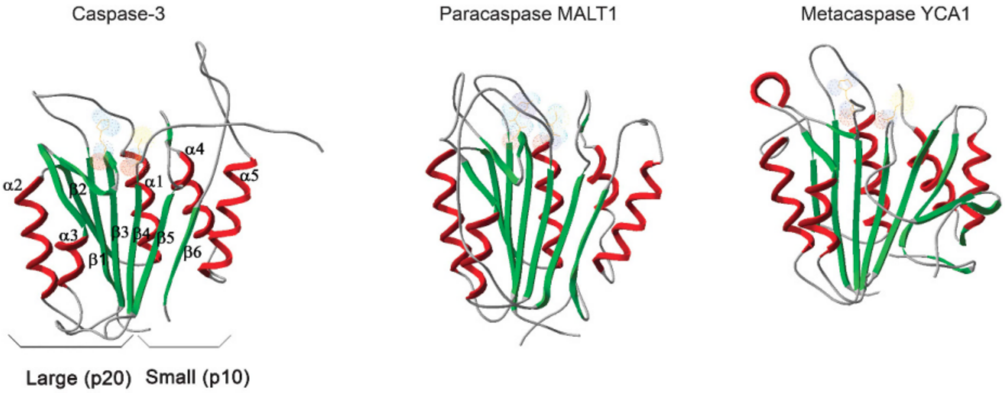


***Figure 1.2.*** *Highly conserved α,β,α-sandwich fold characteristic of the clan CD of cysteine proteases. Adapted from* [84]

In terms of reactivity they also share some characteristics. On the one hand, the specificity is dominated only by the residue at position P1. Namely, in the C14 family hydrolysis is performed after an Asp, in C13 after an Asn, in C11 and C50 after an Arg, and in C25 after a Lys or an Arg depending on the protease. And on the other hand, none of the proteases of the clan are inhibited by E-64,[85] a known inhibitor of cysteine proteases.[62]

Among the eight families that compose the clan CD, one of the smallest is the family C25. It is composed of two proteases produced by the pathogen *Porphyromonas gingivalis*. These two proteases, also known as gingipains (gp), are multidomain proteases and most of them have a hemagglutinin domain at the C-terminal end. The family shows high specificity for Lys residues in the case of lysisne gp (Kgp) and Arg in the case of arginine gingipain (Rgp). Although their main biological activity is associated with the progression of *P. gingivalis* parasitization, causing gingivitis, they have been also widely associated with the development of other medical conditions in humans.[62]

#### 1.2.1.2 ARGININE GINGIPAIN B

The hydrolytic activity of all arginine gingipain variants is strictly limited to the recognition of the arginine residue at the P1 position. None of them shows any particular selectivity or enhancement of activity towards any other residue at positions neighboring P1.[86–89]This selectivity has been extensively tested by incubating Rgp's with different proteins, e.g. glucagon, α and β-globin, ribonucleases and lysozymes. The cleavage products have been evaluated and it has been concluded that it responds only to the presence of Arg and not Lys or other amino acids and that it cuts in all of them. As a consequence of this specific but broad activity, Rgp's degrade a wide range of protein constituents in human connective tissue, cell membrane proteins and receptors, cytokines and plasma proteins, heme and iron-binding proteins, immunoglobulins and peptide regulators.[90,91] This, together with the fact that they are produced in large quantities, strongly impact and promote the parasitic activity of the opportunistic coccobacillus *P. gingivalis.*

The catalytic activity of Rgp's is, in all cases, dependent on the presence of reducing agents.[91] As for the whole clan, they show little impact on the activity in the presence of the inhibitor E-64.[85] However, they are strongly inhibited by peptidyl chloromethanes with Arg moieties at the P1 position. This makes these compounds ideal for the titration of Rgp's.[92] In fact, it is in complex with a peptidyl chloromethane that the crystal structure[83] that was employed in this study as initial coordinates was reported. Furthermore, Rgp's unlike K gingipains are inhibited in the presence of $Zn^{2+}$.[92]

Rgp's occur as 110 or 95 proteins for those products of the *rgpA* gene and 70-90 or 50 kDa for the translation products of the *rgpB* gene.[93–96] Those encoded by the *rgpA* gene, known as HRgpA are composed of a catalytic unit (50 kDa) in non-covalent association with polypeptides derived from the long (985 residues-long) C-terminal tail of the original translation product of the gene. The *rgpB* gene encodes for a 50 kDa enzyme (RgpB) quite similar to the catalytic domain of that encoded by *rgpA* but flanked by a longer fragment at the N-terminal and without the C-terminal peptide fragment with hemagglutinin/adhesin activity. The initial structure of RgpB consists of 435 amino acids. Both HRgpA and RgpB variants can associate with a glycoprotein to give way to their membrane-type mt-HRgpA and mt-RgpB versions. All R gingipains show a low pI value, close to 4.0. RgpB is more stable than the average of cystein proteases and shows activity at high concentrations of denaturing agents such as urea, SDS, Triton X-100 or decylpyranoside. The optimal pH for the catalytic activity of HRgpA is 7.5, whereas RgpB shows a wide and flat range of activity from 6.5 (50%) to 9.5 (100%), beyond which limits the activity decays sharply due to enzymatic instability.[62,91,97] These characteristics, together with the high selectivity for arginine residues, makes RgpB as dangerous to health as it is useful as a tool in protein chemistry.

**Ternary structure.** The best solved crystal structure for RgpB is the one identified with the PDB-ID code: 1CVR,[83] with a resolution of 2 Å and which was the base structure for our simulations. It has a 435-residue, single-poly-peptide chain structure, which can be divided into: a) the catalytic domain (1-351) and an immunoglobulin-like domain (352-435). The catalytic domain can be divided into two subdomains A and B, similar to the subdomains of the catalytic domain of caspase-1. Subdomain B contains the catalytic residues His211 and Cys 244, and the two residues responsible for selectivity toward arginine residues, namely Trp284 and Asp163. These two residues interact with the guanidinium group of arginine in the substrate by salt bridging with Asp163 and by π-stacking with Trp284. The combination of hydrogen bonds with the negatively charged aspartate and π-stacking with tryptophan is a pattern that is impossible to replicate with other amino acids and it is thought that they are the key to the high selectivity of RgpB. Glu152, on the other hand, plays the role of a residue that stabilizes and orients the catalytic His.

#### 1.2.1.3 ARGININE GINGIPAIN B IN ALZHEIMER'S DISEASE

As mentioned above, the main role of RgpB is to promote the pathogenic activity of *P. gingivalis.* That is, its activity is primarily associated with the development of oropharyngeal problems. However, it has been detected in patients with other pathologies, associating RgpB with them. The range of pathologies goes from placental problems, atherosclerosis, problems in femoral and coronary arteries, and neurodegenerative conditions.[98–101] One association, which is of particular interest because of its pharmacological potential, is that between the brain load of gingipains and the progression of Alzheimer's disease, one of the most challenging medical issues nowadays.[102,103] A study published in 2019 by S. S. Dominy and coworkers concluded that the progression of Alzheimer's disease is associated with the presence of gingipains proteases in the brain and that inhibition of these with small molecules has neuroprotective properties.[104] The key points of the study are listed and briefly explained below:

a) *Diagnosis of Alzheimer's disease correlates with gingipain levels in brain:* immunohistochemical (IHC) studies were performed on microarrays of nucleus cerebral tissues from Alzheimer's disease patients and healthy individuals. Specific antibodies to gingipain R and gingipain K were used. These showed that 96% of the diseased patients had the presence of RgpB and 91% of Kgp in the brain tissue. Although it is usual to find them in healthy people as well, it was shown that there is a correlation between the protease load in the brain and the dementia status of the patient, being the load much higher in patients than in healthy people. Other experimental methods were performed to confirm the correlations.

b) *RgpB colocalizes with neurons, astrocytes and pathology in Alzheimer's disease patients' hippocampus:* patients with Alzheimer's disease show rapid deterioration in hippocampus brain areas. By labeling with RgpB antibodies and bright-field microscopy, the presence of RgpB was confirmed in the dentate gyrus and CA3, CA2 and CA1 neurons of the hippocampus of affected patients. Using immunofluorescence, RgpB colocalized mainly with neurons, as well as with occasional astrocytes, but not with microglia (Iba1). Thus confirming the presence of R gingipains in Alzheimer's disease-affected neurons and not in others.

c) *The Tau protein is cleaved by gingipains:* the impact of tau protein tangles in Alzheimer's disease is well known. Therefore, S. S. Dominy and colleagues decided to see if tau was a target of gingipain proteolysis. It has been proposed that tau truncation and fragmentation play a key role in inducing the formation of insoluble and hyperphosphorylated tau in AD. As a result of their analysis, they concluded that there was a dose-dependent decrease in the amount of soluble tau protein in cells infected with *P. gingivalis*, whereas those cells infected with *P. gingivalis*-gingipain mutants and uninfected cells showed stable levels of tau solubility. Additionally, by mass spectrometry they were able to identify that the tau protein fragments in the insoluble plaques came from cleavages made after arginine and lysine residues, congruent with the activity of gingipain proteases.

d) *Gingipain inhibitors are neuroprotective:* S. S. Dominy and coworkers determined that pretreatment of gingipains with iodoacetamide, an irreversible inhibitor of cysteine protease, prevented gingipain-induced plaques aggregation. This indicated that the proteolytic activity of gingipains was the cause of the morphological changes. So the next step was to determine whether inhibiting these prevented neurological damage. What they found, by infecting the brains of mice with doses of gingipains, was that the gingipain-injected mice had significantly more degenerated neurons

> than the saline-injected control mice. However, neurodegeneration was blocked in a dose-dependent manner by applying gingipain inhibitor treatments for days. For this, they carried out the *in silico* design of potent irreversible inhibitors (inhibitory concentration $IC_{50} < 50$ pM) for both Kgp and RgpB. These were subsequently reported in patents PCT/US2015/054050 and PCT/US2016/061197.

Dominy and coworkers' study, partialy reported in the patents, and the advanced clinical trials indicates that these inhibitors have reached as potential treatment for neurodegeneration, put gingipains in the spotlight of Alzheimer's disease research.

## 1.3 REFERENCES

(1) Bugg, T. D. H. Introduction to Enzyme and Coenzyme Chemistry: Third Edition. *Introd. to Enzym. Coenzyme Chem. Third Ed.* **2012**. https://doi.org/10.1002/9781118348970.

(2) Bairoch, A.; Apweiler, R. The SWISS-PROT Protein Sequence Data Bank and Its Supplement TrEMBL. *Nucleic Acids Res.* **1997**, *25* (1), 31–36. https://doi.org/10.1093/nar/25.1.31.

(3) Bateman, A.; Martin, M.-J.; Orchard, S.; Magrane, M.; Agivetova, R.; Ahmad, S.; Alpi, E.; Bowler-Barnett, E. H.; Britto, R.; Bursteinas, B.; Bye-A-Jee, H.; Coetzee, R.; Cukura, A.; Da Silva, A.; Denny, P.; Dogan, T.; Ebenezer, T.; Fan, J.; Castro, L. G.; Garmiri, P.; Georghiou, G.; Gonzales, L.; Hatton-Ellis, E.; Hussein, A.; Ignatchenko, A.; Insana, G.; Ishtiaq, R.; Jokinen, P.; Joshi, V.; Jyothi, D.; Lock, A.; Lopez, R.; Luciani, A.; Luo, J.; Lussi, Y.; MacDougall, A.; Madeira, F.; Mahmoudy, M.; Menchi, M.; Mishra, A.; Moulang, K.; Nightingale, A.; Oliveira, C. S.; Pundir, S.; Qi, G.; Raj, S.; Rice, D.; Lopez, M. R.; Saidi, R.; Sampson, J.; Sawford, T.; Speretta, E.; Turner, E.; Tyagi, N.; Vasudev, P.; Volynkin, V.; Warner, K.; Watkins, X.; Zaru, R.; Zellner, H.; Bridge, A.; Poux, S.; Redaschi, N.; Aimo, L.; Argoud-Puy, G.; Auchincloss, A.; Axelsen, K.; Bansal, P.; Baratin, D.; Blatter, M.-C.; Bolleman, J.; Boutet, E.; Breuza, L.; Casals-Casas, C.; de Castro, E.; Echioukh, K. C.; Coudert, E.; Cuche, B.; Doche, M.; Dornevil, D.; Estreicher, A.; Famiglietti, M. L.; Feuermann, M.; Gasteiger, E.; Gehant, S.; Gerritsen, V.; Gos, A.; Gruaz-Gumowski, N.; Hinz, U.; Hulo, C.; Hyka-Nouspikel, N.; Jungo, F.; Keller, G.; Kerhornou, A.; Lara, V.; Le Mercier, P.; Lieberherr, D.; Lombardot, T.; Martin, X.; Masson, P.; Morgat, A.; Neto, T. B.; Paesano, S.; Pedruzzi, I.; Pilbout, S.; Pourcel, L.; Pozzato, M.; Pruess, M.; Rivoire, C.; Sigrist, C.; Sonesson, K.; Stutz, A.; Sundaram, S.; Tognolli, M.; Verbregue, L.; Wu, C. H.; Arighi, C. N.; Arminski, L.; Chen, C.; Chen, Y.; Garavelli, J. S.; Huang, H.; Laiho, K.; McGarvey, P.; Natale, D. A.; Ross, K.; Vinayaka, C. R.; Wang, Q.; Wang, Y.; Yeh, L.-S.; Zhang, J.; Ruch, P.;

Teodoro, D. UniProt: The Universal Protein Knowledgebase in 2021. *Nucleic Acids Res.* **2021**, *49* (D1), D480–D489. https://doi.org/10.1093/nar/gkaa1100.

(4) Babbi, G.; Baldazzi, D.; Savojardo, C.; Martelli, P. L.; Casadio, R. Highlighting Human Enzymes Active in Different Metabolic Pathways and Diseases: The Case Study of EC 1.2.3.1 and EC 2.3.1.9. *Biomedicines* **2020**, *8* (8), 250. https://doi.org/10.3390/biomedicines8080250.

(5) Purich, D. L. Enzyme Kinetics: Catalysis and Control a Reference of Theory and Best-Practice Methods. *Enzym. Kinet. Catal. Control A Ref. Theory Best-Practice Methods* **2010**, 1–892. https://doi.org/10.1016/C2009-0-61154-5.

(6) Robertson, J. G. Enzymes as a Special Class of Therapeutic Target: Clinical Drugs and Modes of Action. *Curr. Opin. Struct. Biol.* **2007**, *17* (6), 674–679. https://doi.org/10.1016/j.sbi.2007.08.008.

(7) Warshel, A.; Sharma, P. K.; Kato, M.; Parson, W. W. Modeling Electrostatic Effects in Proteins. *Biochim. Biophys. Acta - Proteins Proteomics* **2006**, *1764* (11), 1647–1676. https://doi.org/10.1016/j.bbapap.2006.08.007.

(8) Wolfenden, R. Transition State Analogues for Enzyme Catalysis. *Nature* **1969**, *223* (5207), 704–705. https://doi.org/10.1038/223704a0.

(9) Pauling, L. Molecular Architecture and Biological Reactions. *Chem. Eng. News Arch.* **1946**, *24* (10), 1375–1377. https://doi.org/10.1021/cen-v024n010.p1375.

(10) Warshel, A. Electrostatic Origin of the Catalytic Power of Enzymes and the Role of Preorganized Active Sites. *J. Biol. Chem.* **1998**, *273* (42), 27035–27038. https://doi.org/10.1074/jbc.273.42.27035.

(11) Schramm, V. L. Enzymatic Transition State Poise and Transition State Analogues. *Acc. Chem. Res.* **2003**, *36* (8), 588–596. https://doi.org/10.1021/ar0200495.

(12) Martí, S.; Roca, M.; Andrés, J.; Moliner, V.; Silla, E.; Tuñón, I.; Bertrán, J. Theoretical Insights in Enzyme Catalysis. *Chem. Soc. Rev.* **2004**, *33* (2), 98–107. https://doi.org/10.1039/B301875J.

(13) Roca, M.; Martí, S.; Andrés, J.; Moliner, V.; Tuñón, I.; Bertrán, J.; Williams, I. H. Theoretical Modeling of Enzyme Catalytic Power: Analysis of "Cratic" and Electrostatic Factors in Catechol O -Methyltransferase. *J. Am. Chem. Soc.* **2003**, *125* (25), 7726–7737. https://doi.org/10.1021/ja0299497.

(14) Luk, L. Y. P.; Javier Ruiz-Pernía, J.; Dawson, W. M.; Roca, M.; Loveridge, E. J.; Glowacki, D. R.; Harvey, J. N.; Mulholland, A. J.; Tuñón, I.; Moliner, V.; Allemann, R. K. Unraveling the Role of Protein Dynamics in Dihydrofolate Reductase Catalysis. *Proc. Natl. Acad. Sci.* **2013**, *110* (41), 16344–16349. https://doi.org/10.1073/pnas.1312437110.

(15) Jencks, W. P. Binding Energy, Specificity, and Enzymic Catalysis: The Circe Effect. In *Advances in Enzymology - and Related Areas of Molecular Biology*; 2006; pp 219–410. https://doi.org/10.1002/9780470122884.ch4.

(16) Villà, J.; Štrajbl, M.; Glennon, T. M.; Sham, Y. Y.; Chu, Z. T.; Warshel, A. How Important Are Entropic Contributions to Enzyme Catalysis? *Proc. Natl. Acad. Sci.* **2000**, *97* (22), 11899–11904. https://doi.org/10.1073/pnas.97.22.11899.

(17) Scott, W. G. Ribozyme Catalysis via Orbital Steering11Edited by J. Doudna. *J. Mol. Biol.* **2001**, *311* (5), 989–999. https://doi.org/10.1006/jmbi.2001.4886.

(18) Hackney, D. D. Binding Energy and Catalysis; 1990; pp 1–36. https://doi.org/10.1016/S1874-6047(08)60194-8.

(19) Mesecar, A. D.; Stoddard, B. L.; Koshland, D. E. Orbital Steering in the Catalytic Power of Enzymes: Small Structural Changes with Large Catalytic Consequences. *Science.* **1997**, *277* (5323), 202–206. https://doi.org/10.1126/science.277.5323.202.

(20) Woodward, R. B.; Hoffmann, R. The Conservation of Orbital Symmetry. *Angew. Chemie Int. Ed. English* **1969**, *8* (11), 781–853. https://doi.org/10.1002/anie.196907811.

(21) Storm, D. R.; Koshland, D. E. A Source for the Special Catalytic Power of Enzymes: Orbital Steering. *Proc. Natl. Acad. Sci.* **1970**, *66* (2), 445–452. https://doi.org/10.1073/pnas.66.2.445.

(22) Bruice, T. C. A View at the Millennium: The Efficiency of Enzymatic Catalysis. *Acc. Chem. Res.* **2002**, *35* (3), 139–148. https://doi.org/10.1021/ar0001665.

(23) Hur, S.; Bruice, T. C. Just a Near Attack Conformer for Catalysis (Chorismate to Prephenate Rearrangements in Water, Antibody, Enzymes, and Their Mutants). *J. Am. Chem. Soc.* **2003**, *125* (35), 10540–10542. https://doi.org/10.1021/ja0357846.

(24) Hur, S.; Bruice, T. C. The near Attack Conformation Approach to the Study of the Chorismate to Prephenate Reaction. *Proc. Natl. Acad. Sci.* **2003**, *100* (21), 12015–12020. https://doi.org/10.1073/pnas.1534873100.

(25) Jencks, W. P. Mechanism of Enzyme Action. *Annu. Rev. Biochem.* **1963**, *32* (1), 639–676. https://doi.org/10.1146/annurev.bi.32.070163.003231.

(26) William P. Jencks. *Catalysis in Chemistry and Enzymology*, First.; McGraw-Hill: New York, 1969.

(27) William P. Jencks. *Catalysis in Chemistry and Enzymology*; Dover Publications, 1987.

(28) Hammes-Schiffer, S. Impact of Enzyme Motion on Activity. *Biochemistry* **2002**, *41* (45), 13335–13343. https://doi.org/10.1021/bi0267137.

(29) Radisky, E. S.; Koshland, D. E. A Clogged Gutter Mechanism for Protease Inhibitors. *Proc. Natl. Acad. Sci.* **2002**, *99* (16), 10316–10321. https://doi.org/10.1073/pnas.112332899.

(30) Benkovic, S. J.; Hammes, G. G.; Hammes-Schiffer, S. Free-Energy Landscape of Enzyme Catalysis. *Biochemistry* **2008**, *47* (11), 3317–3321. https://doi.org/10.1021/bi800049z.

(31) Eisenmesser, E. Z.; Millet, O.; Labeikovsky, W.; Korzhnev, D. M.; Wolf-Watz, M.; Bosco, D. A.; Skalicky, J. J.; Kay, L. E.; Kern, D. Intrinsic Dynamics of an Enzyme Underlies Catalysis. *Nature* **2005**, *438* (7064), 117–121. https://doi.org/10.1038/nature04105.

(32) Hay, S.; Scrutton, N. S. Good Vibrations in Enzyme-Catalysed Reactions. *Nat. Chem.* **2012**, *4* (3), 161–168. https://doi.org/10.1038/nchem.1223.

(33) Bhabha, G.; Lee, J.; Ekiert, D. C.; Gam, J.; Wilson, I. A.; Dyson, H. J.; Benkovic, S. J.; Wright, P. E. A Dynamic Knockout Reveals That Conformational Fluctuations Influence the Chemical Step of Enzyme Catalysis. *Science.* **2011**, *332* (6026), 234–238. https://doi.org/10.1126/science.1198542.

(34) Michaelis, L.; Menten, M. L. Die Kinetik Der Invertinwirkung. *Biochem. Z.* **1913**, *49*, 333–369.

(35) Johnson, K. A.; Goody, R. S. The Original Michaelis Constant: Translation of the 1913 Michaelis–Menten Paper. *Biochemistry* **2011**, *50* (39), 8264–8269. https://doi.org/10.1021/bi201284u.

(36) *The IUPAC Compendium of Chemical Terminology*; McNaught, A. D., Wilkinson, A., Eds.; International Union of Pure and Applied Chemistry (IUPAC): Research Triangle Park, NC, 2019. https://doi.org/10.1351/goldbook.

(37) Truhlar, D. G.; Hase, W. L.; Hynes, J. T. Current Status of Transition-State Theory. *J. Phys. Chem.* **1983**, *87* (26), 5523–5523. https://doi.org/10.1021/j150644a044.

(38) Baillie, T. A. Targeted Covalent Inhibitors for Drug Design. *Angew. Chemie Int. Ed.* **2016**, *55* (43), 13408–13421. https://doi.org/10.1002/anie.201601091.

(39) Gehringer, M.; Laufer, S. A. Emerging and Re-Emerging Warheads for Targeted Covalent Inhibitors: Applications in Medicinal Chemistry and Chemical Biology. *J. Med. Chem.* **2019**, *62* (12), 5673–5724. https://doi.org/10.1021/acs.jmedchem.8b01153.

(40) Tuley, A.; Fast, W. The Taxonomy of Covalent Inhibitors. *Biochemistry* **2018**, *57* (24), 3326–3337. https://doi.org/10.1021/acs.biochem.8b00315.

(41) Coleman, C. I.; Limone, B.; Sobieraj, D. M.; Lee, S.; Roberts, M. S.; Kaur, R.; Alam, T. Dosing Frequency and Medication Adherence in Chronic Disease. *J. Manag. Care Pharm.* **2012**, *18* (7), 527–539. https://doi.org/10.18553/jmcp.2012.18.7.527.

(42) Walter, A. O.; Sjin, R. T. T.; Haringsma, H. J.; Ohashi, K.; Sun, J.; Lee, K.; Dubrovskiy, A.; Labenski, M.; Zhu, Z.; Wang, Z.; Sheets, M.; St Martin, T.; Karp, R.; van Kalken, D.; Chaturvedi, P.; Niu, D.; Nacht, M.; Petter, R. C.; Westlin, W.; Lin, K.; Jaw-Tsai, S.; Raponi, M.; Van Dyke, T.; Etter, J.; Weaver, Z.; Pao, W.; Singh, J.; Simmons, A. D.; Harding, T. C.; Allen, A. Discovery of a Mutant-Selective Covalent Inhibitor of EGFR That Overcomes T790M-Mediated Resistance in NSCLC. *Cancer Discov.* **2013**, *3* (12), 1404–1415. https://doi.org/10.1158/2159-8290.CD-13-0314.

(43) Kempson, J.; Ovalle, D.; Guo, J.; Wrobleski, S. T.; Lin, S.; Spergel, S. H.; Duan, J. J.-W.; Jiang, B.; Lu, Z.; Das, J.; Yang, B. V.; Hynes, J.; Wu, H.; Tokarski, J.; Sack, J. S.; Khan, J.; Schieven, G.; Blatt, Y.; Chaudhry, C.; Salter-Cid, L. M.; Fura, A.; Barrish, J. C.; Carter, P. H.; Pitts, W. J. Discovery of Highly Potent, Selective, Covalent Inhibitors of JAK3. *Bioorg.*

*Med. Chem. Lett.* **2017**, *27* (20), 4622–4625. https://doi.org/10.1016/j.bmcl.2017.09.023.

(44) Morrison, J. F.; Walsh, C. T. The Behavior and Significance of Slow-Binding Enzyme Inhibitors; 2006; pp 201–301. https://doi.org/10.1002/9780470123072.ch5.

(45) So, W. H.; Zhang, Y.; Kang, W.; Wong, C. T. T.; Sun, H.; Xia, J. Site-Selective Covalent Reactions on Proteinogenic Amino Acids. *Curr. Opin. Biotechnol.* **2017**, *48*, 220–227. https://doi.org/10.1016/j.copbio.2017.06.003.

(46) deGruyter, J. N.; Malins, L. R.; Baran, P. S. Residue-Specific Peptide Modification: A Chemist's Guide. *Biochemistry* **2017**, *56* (30), 3863–3873. https://doi.org/10.1021/acs.biochem.7b00536.

(47) Shannon, D. A.; Weerapana, E. Covalent Protein Modification: The Current Landscape of Residue-Specific Electrophiles. *Curr. Opin. Chem. Biol.* **2015**, *24*, 18–26. https://doi.org/10.1016/j.cbpa.2014.10.021.

(48) Baslé, E.; Joubert, N.; Pucheault, M. Protein Chemical Modification on Endogenous Amino Acids. *Chem. Biol.* **2010**, *17* (3), 213–227. https://doi.org/10.1016/j.chembiol.2010.02.008.

(49) Plapp, B. V. [25] Application of Affinity Labeling for Studying Structure and Function of Enzymes; 1982; pp 469–499. https://doi.org/10.1016/S0076-6879(82)87027-4.

(50) Lanning, B. R.; Whitby, L. R.; Dix, M. M.; Douhan, J.; Gilbert, A. M.; Hett, E. C.; Johnson, T. O.; Joslyn, C.; Kath, J. C.; Niessen, S.; Roberts, L. R.; Schnute, M. E.; Wang, C.; Hulce, J. J.; Wei, B.; Whiteley, L. O.; Hayward, M. M.; Cravatt, B. F. A Road Map to Evaluate the Proteome-Wide Selectivity of Covalent Kinase Inhibitors. *Nat. Chem. Biol.* **2014**, *10* (9), 760–767. https://doi.org/10.1038/nchembio.1582.

(51) Schwartz, P. A.; Kuzmic, P.; Solowiej, J.; Bergqvist, S.; Bolanos, B.; Almaden, C.; Nagata, A.; Ryan, K.; Feng, J.; Dalvie, D.; Kath, J. C.; Xu,

M.; Wani, R.; Murray, B. W. Covalent EGFR Inhibitor Analysis Reveals Importance of Reversible Interactions to Potency and Mechanisms of Drug Resistance. *Proc. Natl. Acad. Sci.* **2014**, *111* (1), 173–178. https://doi.org/10.1073/pnas.1313733111.

(52) Solca, F.; Dahl, G.; Zoephel, A.; Bader, G.; Sanderson, M.; Klein, C.; Kraemer, O.; Himmelsbach, F.; Haaksma, E.; Adolf, G. R. Target Binding Properties and Cellular Activity of Afatinib (BIBW 2992), an Irreversible ErbB Family Blocker. *J. Pharmacol. Exp. Ther.* **2012**, *343* (2), 342–350. https://doi.org/10.1124/jpet.112.197756.

(53) Krantz, A. A Classification of Enzyme Inhibitors. *Bioorg. Med. Chem. Lett.* **1992**, *2* (11), 1327–1334. https://doi.org/10.1016/S0960-894X(00)80507-7.

(54) Silverman, R. B. [10] Mechanism-Based Enzyme Inactivators; 1995; pp 240–283. https://doi.org/10.1016/0076-6879(95)49038-8.

(55) López-Otín, C.; Bond, J. S. Proteases: Multifunctional Enzymes in Life and Disease. *J. Biol. Chem.* **2008**, *283* (45), 30433–30437. https://doi.org/10.1074/jbc.R800035200.

(56) Neurath, H.; Walsh, K. A. Role of Proteolytic Enzymes in Biological Regulation (a Review). *Proc. Natl. Acad. Sci.* **1976**, *73* (11), 3825–3832. https://doi.org/10.1073/pnas.73.11.3825.

(57) López-Otín, C.; Overall, C. M. Protease Degradomics: A New Challenge for Proteomics. *Nat. Rev. Mol. Cell Biol.* **2002**, *3* (7), 509–519. https://doi.org/10.1038/nrm858.

(58) Ehrmann, M.; Clausen, T. Proteolysis as a Regulatory Mechanism. *Annu. Rev. Genet.* **2004**, *38* (1), 709–724. https://doi.org/10.1146/annurev.genet.38.072902.093416.

(59) Sauer, R. T.; Bolon, D. N.; Burton, B. M.; Burton, R. E.; Flynn, J. M.; Grant, R. A.; Hersch, G. L.; Joshi, S. A.; Kenniston, J. A.; Levchenko, I.; Neher, S. B.; Oakes, E. S. C.; Siddiqui, S. M.; Wah, D. A.; Baker, T. A.

Sculpting the Proteome with AAA+ Proteases and Disassembly Machines. *Cell* **2004**, *119* (1), 9–18. https://doi.org/10.1016/j.cell.2004.09.020.

(60) Siegel, R. M. Caspases at the Crossroads of Immune-Cell Life and Death. *Nat. Rev. Immunol.* **2006**, *6* (4), 308–317. https://doi.org/10.1038/nri1809.

(61) Pérez-Silva, J. G.; Español, Y.; Velasco, G.; Quesada, V. The Degradome Database: Expanding Roles of Mammalian Proteases in Life and Disease. *Nucleic Acids Res.* **2016**, *44* (D1), D351–D355. https://doi.org/10.1093/nar/gkv1201.

(62) Barrett, A. J.; Rawlings, N. D.; Woessner, J. F. Handbook of Proteolytic Enzymes: Second Edition. *Handb. Proteolytic Enzym. Second Ed.* **2004**, *1*, 1–1140. https://doi.org/10.1016/C2009-0-03628-9.

(63) Glazer, A. N.; Smith, E. L. 14 Papain and Other Plant Sulfhydryl Proteolytic Enzymes; 1971; pp 501–546. https://doi.org/10.1016/S1874-6047(08)60405-9.

(64) Whitaker, J. R.; Bender, M. L. Kinetics of Papain-Catalyzed Hydrolysis of α-N-Benzoyl-L-Arginine Ethyl Ester and α-N-Benzoyl-L-Argininamide 1. *J. Am. Chem. Soc.* **1965**, *87* (12), 2728–2737. https://doi.org/10.1021/ja01090a034.

(65) Lucas, E. C.; Williams, A. PH Dependencies of Individual Rate Constants in Papain-Catalyzed Reactions. *Biochemistry* **1969**, *8* (12), 5125–5135. https://doi.org/10.1021/bi00840a067.

(66) Allen, K. G. D.; Stewart, J. A.; Johnson, P. E.; Wettlaufer, D. G. Identification of the Functional Ionic Groups of Papain by PH/Rate Profile Analysis. *Eur. J. Biochem.* **1978**, *87* (3), 575–582. https://doi.org/10.1111/j.1432-1033.1978.tb12409.x.

(67) Drenth, J.; Kalk, K. H.; Swen, H. M. Binding of Chloromethyl Ketone Substrate Analogs to Crystalline Papain. *Biochemistry* **1976**, *15* (17), 3731–3738. https://doi.org/10.1021/bi00662a014.

(68) Brocklehurst, K.; Little, G. Reactivities of the Various Protonic States in the Reactions of Papain and of L-Cysteine with 2,2′- and with 4,4′- Dipyridyl Disulphide: Evidence for Nucleophilic Reactivity in the Un-Ionized Thiol Group of the Cysteine-25 Residue of Papain Occasioned By. *Biochem. J.* **1972**, *128* (2), 471–474. https://doi.org/10.1042/bj1280471.

(69) Polgar, L. On the Mode of Activation of the Catalytically Essential Sulfhydryl Group of Papain. *Eur. J. Biochem.* **1973**, *33* (1), 104–109. https://doi.org/10.1111/j.1432-1033.1973.tb02660.x.

(70) Asboth, B.; Stokum, E.; Khan, I. U.; Polgar, L. Mechanism of Action of Cysteine Proteinases: Oxyanion Binding Site Is Not Essential in the Hydrolysis of Specific Substrates. *Biochemistry* **1985**, *24* (3), 606–609. https://doi.org/10.1021/bi00324a010.

(71) Asboth, B.; Polgar, L. Transition-State Stabilization at the Oxyanion Binding Sites of Serine and Thiol Proteinases: Hydrolyses of Thiono and Oxygen Esters. *Biochemistry* **1983**, *22* (1), 117–122. https://doi.org/10.1021/bi00270a017.

(72) Menard, R.; Carriere, J.; Laflamme, P.; Plouffe, C.; Khouri, H. E.; Vernet, T.; Tessier, D. C.; Thomas, D. Y.; Storer, A. C. Contribution of the Glutamine 19 Side Chain to Transition-State Stabilization in the Oxyanion Hole of Papain. *Biochemistry* **1991**, *30* (37), 8924–8928. https://doi.org/10.1021/bi00101a002.

(73) Menard, R.; Plouffe, C.; Laflamme, P.; Vernet, T.; Tessier, D. C.; Thomas, D. Y.; Storer, A. C. Modification of the Electrostatic Environment Is Tolerated in the Oxyanion Hole of the Cysteine Protease Papain. *Biochemistry* **1995**, *34* (2), 464–471. https://doi.org/10.1021/bi00002a010.

(74) Menard, R.; Khouri, H. E.; Plouffe, C.; Laflamme, P.; Dupras, R.; Vernet, T.; Tessier, D. C.; Thomas, D. Y.; Storer, A. C. Importance of Hydrogen-Bonding Interactions Involving the Side Chain of Asp158 in the Catalytic

Mechanism of Papain. *Biochemistry* **1991**, *30* (22), 5531–5538. https://doi.org/10.1021/bi00236a028.

(75) Watts, A.; Haffez, A.; Gul, S.; Verma, C.; Thomas, E. W.; Brocklehurst, K. Effects of Site-Specific Mutations on the Kinetically Influential Ionizations of Papain. *Biochem. Soc. Trans.* **1997**, *25* (1), 84S-84S. https://doi.org/10.1042/bst025084s.

(76) Taylor, M. A. J.; Baker, K. C.; Connerton, I. F.; Cummings, N. J.; Harris, G. W.; Henderson, I. M. J.; Jones, S. T.; Pickersgill, R. W.; Sumner, I. G.; Warwicker, J.; Goodenough, P. W. An Unequivocal Example of Cysteine Proteinase Activity Affected by Multiple Electrostatic Interactions. *"Protein Eng. Des. Sel.* **1994**, *7* (10), 1267–1276. https://doi.org/10.1093/protein/7.10.1267.

(77) Noble, M. A.; Gul, S.; Verma, C. S.; Brocklehurst, K. Ionization Characteristics and Chemical Influences of Aspartic Acid Residue 158 of Papain and Caricain Determined by Structure-Related Kinetic and Computational Techniques: Multiple Electrostatic Modulators of Active-Centre Chemistry. *Biochem. J.* **2000**, *351* (3), 723–733. https://doi.org/10.1042/bj3510723.

(78) Pinitglang, S.; Watts, A. B.; Patel, M.; Reid, J. D.; Noble, M. A.; Gul, S.; Bokth, A.; Naeem, A.; Patel, H.; Thomas, E. W.; Sreedharan, S. K.; Verma, C.; Brocklehurst, K. A Classical Enzyme Active Center Motif Lacks Catalytic Competence until Modulated Electrostatically. *Biochemistry* **1997**, *36* (33), 9968–9982. https://doi.org/10.1021/bi9705974.

(79) Chen, J.-M.; Rawlings, N. D.; Stevens, R. A. E.; Barrett, A. J. Identification of the Active Site of Legumain Links It to Caspases, Clostripain and Gingipains in a New Clan of Cysteine Endopeptidases. *FEBS Lett.* **1998**, *441* (3), 361–365. https://doi.org/10.1016/S0014-5793(98)01574-9.

(80) Walker, N. P. C.; Talanian, R. V.; Brady, K. D.; Dang, L. C.; Bump, N. J.; Ferenza, C. R.; Franklin, S.; Ghayur, T.; Hackett, M. C.; Hammill, L. D.;

Herzog, L.; Hugunin, M.; Houy, W.; Mankovich, J. A.; McGuiness, L.; Orlewicz, E.; Paskind, M.; Pratt, C. A.; Reis, P.; Summani, A.; Terranova, M.; Welch, J. P.; Xiong, L.; Möller, A.; Tracey, D. E.; Kamen, R.; Wong, W. W. Crystal Structure of the Cysteine Protease Interleukin-1β-Converting Enzyme: A (P20/P10)2 Homodimer. *Cell* **1994**, *78* (2), 343–352. https://doi.org/10.1016/0092-8674(94)90303-4.

(81) Wilson, K. P.; Black, J.-A. F.; Thomson, J. A.; Kim, E. E.; Griffith, J. P.; Navia, M. A.; Murcko, M. A.; Chambers, S. P.; Aldape, R. A.; Raybuck, S. A.; Livingston, D. J. Structure and Mechanism of Interleukin-Lβ Converting Enzyme. *Nature* **1994**, *370* (6487), 270–275. https://doi.org/10.1038/370270a0.

(82) Timmer, J. C.; Zhu, W.; Pop, C.; Regan, T.; Snipas, S. J.; Eroshkin, A. M.; Riedl, S. J.; Salvesen, G. S. Structural and Kinetic Determinants of Protease Substrates. *Nat. Struct. Mol. Biol.* **2009**, *16* (10), 1101–1108. https://doi.org/10.1038/nsmb.1668.

(83) Eichinger, A.; Beisel, H. G.; Jacob, U.; Huber, R.; Medrano, F. J.; Banbula, A.; Potempa, J.; Travis, J.; Bode, W. Crystal Structure of Gingipain R: An Arg-Specific Bacterial Cysteine Proteinase with a Caspase-like Fold. *EMBO J.* **1999**, *18*, 5453–5462. https://doi.org/10.1093/emboj/18.20.5453.

(84) Minina, E. A.; Coll, N. S.; Tuominen, H.; Bozhkov, P. V. Metacaspases versus Caspases in Development and Cell Fate Regulation. *Cell Death Differ.* **2017**, *24* (8), 1314–1325. https://doi.org/10.1038/cdd.2017.18.

(85) Hanada, K.; Tamai, M.; Yamagishi, M.; Ohmura, S.; Sawada, J.; Tanaka, I. Isolation and Characterization of E–64, a New Thiol Protease Inhibitor. *Agric. Biol. Chem.* **1978**, *42* (3), 523–528. https://doi.org/10.1080/00021369.1978.10863014.

(86) Shah, H. N.; Gharbia, S. E.; Kowlessur, D.; Wilkie, E.; Brocklehurst, K. Isolation and Characterization of Gingivain, a Cysteine Proteinase from

Porphyromonas Gingivalis Strain W83. *Biochem. Soc. Trans.* **1990**, *18* (4), 578–579. https://doi.org/10.1042/BST0180578.

(87) Chen, Z.; Potempa, J.; Polanowski, A.; Wikstrom, M.; Travis, J. Purification and Characterization of a 50-KDa Cysteine Proteinase (Gingipain) from Porphyromonas Gingivalis. *J. Biol. Chem.* **1992**, *267* (26), 18896–18901. https://doi.org/10.1016/S0021-9258(19)37045-0.

(88) Bedi, G. S. Purification and Characterization of Lysine- and Arginine-Specific Gingivain Proteases from Porphyromonas Gingivalis. *Prep. Biochem.* **1994**, *24* (3), 251–261. https://doi.org/10.1080/10826069408010097.

(89) Pike, R.; McGraw, W.; Potempa, J.; Travis, J. Lysine- and Arginine-Specific Proteinases from Porphyromonas Gingivalis. Isolation, Characterization, and Evidence for the Existence of Complexes with Hemagglutinins. *J. Biol. Chem.* **1994**, *269* (1), 406–411. https://doi.org/10.1016/S0021-9258(17)42365-9.

(90) Banbula, A.; Mak, P.; Smoluch, M.; Travis, J.; Potempa, J. Arginine-Specific Cysteine Proteinase from Porphyromonas Gingivalis as a Convenient Tool in Protein Chemistry. *Biol. Chem.* **2001**, *382* (9), 1399–1404. https://doi.org/10.1515/BC.2001.172.

(91) Rangarajan, M.; Smith, S. J. M.; U, S.; Curtis, M. A. Biochemical Characterization of the Arginine-Specific Proteases of Porphyromonas Gingivalis W50 Suggests a Common Precursor. *Biochem. J.* **1997**, *323* (3), 701–709. https://doi.org/10.1042/bj3230701.

(92) Potempa, J.; Pike, R.; Travis, J. Titration and Mapping of the Active Site of Cysteine Proteinases from Porphyromonas Gingivalis (Gingipains) Using Peptidyl Chloromethanes. *Biol. Chem.* **1997**, *378* (3–4), 223–230. https://doi.org/10.1515/BCHM.1997.378.3-4.223/MACHINEREADABLECITATION/RIS.

(93) Aduse-Opoku, J.; Muir, J.; Slaney, J. M.; Rangarajan, M.; Curtis, M. A. Characterization, Genetic Analysis, and Expression of a Protease Antigen (PrpRI) of Porphyromonas Gingivalis W50. *Infect. Immun.* **1995**, *63* (12), 4744–4754. https://doi.org/10.1128/IAI.63.12.4744-4754.1995.

(94) Nakayama, K.; Kadowaki, T.; Okamoto, K.; Yamamoto, K. Construction and Characterization of Arginine-Specific Cysteine Proteinase (Arg-Gingipain)-Deficient Mutants of Porphyromonas Gingivalis: Evidence for Significant Contribution of Arg-Gingipain to Virulence. *J. Biol. Chem.* **1995**, *270* (40), 23619–23626. https://doi.org/10.1074/JBC.270.40.23619.

(95) Okamoto, K.; Misumi, Y.; Kadowaki, T.; Yoneda, M.; Yamamoto, K.; Ikehara, Y. Structural Characterization of Argingipain, a Novel Arginine-Specific Cysteine Proteinase as a Major Periodontal Pathogenic Factor from Porphyromonas Gingivalis. *Arch. Biochem. Biophys.* **1995**, *316* (2), 917–925. https://doi.org/10.1006/ABBI.1995.1123.

(96) Pavloff, N.; Pemberton, P. A.; Potempa, J.; Chen, W. C. A.; Pike, R. N.; Prochazka, V.; Kiefer, M. C.; Travis, J.; Barr, P. J. Molecular Cloning and Characterization of Porphyromonas Gingivalis Lysine-Specific Gingipain: A New Member of an Emerging Family of Pathogenic Bacterial Cysteine Proteinases. *J. Biol. Chem.* **1997**, *272* (3), 1595–1600. https://doi.org/10.1074/JBC.272.3.1595.

(97) Potempa, J.; Mikolajczyk-Pawlinska, J.; Brassell, D.; Nelson, D.; Thøgersen, I. B.; Enghild, J. J.; Travis, J. Comparative Properties of Two Cysteine Proteinases (Gingipains R), the Products of Two Related but Individual Genes of Porphyromonas Gingivalis. *J. Biol. Chem.* **1998**. https://doi.org/10.1074/jbc.273.34.21648.

(98) Mahendra, J.; Mahendra, L.; Kurian, V. M.; Jaishankar, K.; Mythilli, R. Prevalence of Periodontal Pathogens in Coronary Atherosclerotic Plaque of Patients Undergoing Coronary Artery Bypass Graft Surgery. *J. Maxillofac. Oral Surg.* **2009**, *8* (2), 108–113. https://doi.org/10.1007/s12663-009-0028-5.

(99) Katz, J.; Chegini, N.; Shiverick, K. T.; Lamont, R. J. Localization of P. Gingivalis in Preterm Delivery Placenta. *J. Dent. Res.* **2009**, *88* (6), 575–578. https://doi.org/10.1177/0022034509338032.

(100) Ishikawa, M.; Yoshida, K.; Okamura, H.; Ochiai, K.; Takamura, H.; Fujiwara, N.; Ozaki, K. Oral Porphyromonas Gingivalis Translocates to the Liver and Regulates Hepatic Glycogen Synthesis through the Akt/GSK-3β Signaling Pathway. *Biochim. Biophys. Acta - Mol. Basis Dis.* **2013**, *1832* (12), 2035–2043. https://doi.org/10.1016/j.bbadis.2013.07.012.

(101) Mougeot, J.-L. C.; Stevens, C. B.; Paster, B. J.; Brennan, M. T.; Lockhart, P. B.; Mougeot, F. K. B. Porphyromonas Gingivalis Is the Most Abundant Species Detected in Coronary and Femoral Arteries. *J. Oral Microbiol.* **2017**, *9* (1), 1281562. https://doi.org/10.1080/20002297.2017.1281562.

(102) Sharma, P.; Srivastava, P.; Seth, A.; Tripathi, P. N.; Banerjee, A. G.; Shrivastava, S. K. Comprehensive Review of Mechanisms of Pathogenesis Involved in Alzheimer's Disease and Potential Therapeutic Strategies. *Prog. Neurobiol.* **2019**, *174*, 53–89. https://doi.org/10.1016/j.pneurobio.2018.12.006.

(103) De Strooper, B.; Karran, E. The Cellular Phase of Alzheimer's Disease. *Cell* **2016**, *164* (4), 603–615. https://doi.org/10.1016/j.cell.2015.12.056.

(104) Dominy, S. S.; Lynch, C.; Ermini, F.; Benedyk, M.; Marczyk, A.; Konradi, A.; Nguyen, M.; Haditsch, U.; Raha, D.; Griffin, C.; Holsinger, L. J.; Arastu-Kapur, S.; Kaba, S.; Lee, A.; Ryder, M. I.; Potempa, B.; Mydel, P.; Hellvard, A.; Adamowicz, K.; Hasturk, H.; Walker, G. D.; Reynolds, E. C.; Faull, R. L. M.; Curtis, M. A.; Dragunow, M.; Potempa, J. Porphyromonas Gingivalis in Alzheimer's Disease Brains: Evidence for Disease Causation and Treatment with Small-Molecule Inhibitors. *Sci. Adv.* **2019**, *5* (1), eaau3333. https://doi.org/10.1126/sciadv.aau3333.

# CHAPTER 2. OBJECTIVES

The abovementioned issues highlights two critical points. First, RgpB is a molecular target with high pharmaceutical potential in the development of new strategies to treat Alzheimer's disease. Second, a detailed understanding of its mechanism of action and inhibition is necessary to be exploited in such applications. The focus of this thesis is to explore and apply different computational strategies to unravel the atomistic details of the catalytic and inhibition mechanisms of RgpB protease. In silico tools, as will be unpacked in the next chapter, allow us to shed light on phenomena that are unattainable experimentally. The interpretation and understanding of the proteolytic activity of RgpB and already known inhibitors can be used in the design of future drugs against Alzehimer's disease.

Bearing this in mind, the specific objectives that have been raised and addressed in the present study are the following:

a) Revealing the reaction mechanism of proteolysis catalyzed by RgpB.

b) Unraveling and understanding the role of the enzyme in the catalytic process.

c) Analyze the non-covalent interactions that govern the binding processes between a set of known inhibitors and the RgpB active site.

d) To associate the different interaction profiles between inhibitors and RgpB to their relative potency.

e) To elucidate the mechanism of covalent binding between irreversible inhibitors and the RgpB protease.

f) To employ new Quantum Mechanics/Molecular Mechanics techniques such as metadynamics to study the mechanism of catalytic and inhibitory enzymatic processes.

# CHAPTER 3. METHODS

## 3.1 MOLECULAR MODELLING

Molecular modeling is defined as the description of molecular phenomena by means of simplified or idealistic descriptions/models. Nowadays, molecular modeling is increasingly associated with the use of computers that allow the efficient handling of more complex systems or at a deeper level of detail. However, at the most fundamental level, a molecular modeler can make use of physical models of spheres and sticks or even simple paper and pencil 2D representations. It should be noted that the simplicity of a model is not necessarily associated with greater optimality. Indeed, a challenge that molecular modelers face on a daily basis is the choice of descriptors that allow them to address the problem of interest without overflowing in time or complexity. Thus, the ability to balance accuracy and simplicity is a virtue well valued in the field.[1]

In quantum mechanics, the most obvious starting point is undoubtedly the Schrödinger equation,[2,3] which, in its most complete and time-dependent expression is:

$$\left\{-\frac{\hbar^2}{2m}\left(\frac{\partial^2}{\partial x^2}+\frac{\partial^2}{\partial y^2}+\frac{\partial^2}{\partial z^2}\right)+V\right\}\Psi(r,t)=i\hbar\frac{\partial\Psi(r,t)}{\partial t} \quad \text{(Eq. 3.1)}$$

By means of this equation it is possible to describe the behavior of a particle of mass m, with a position vector in r coordinates and under the influence of a potential V. ħ is Planck's constant divided by 2π and Ψ is the wave function describing the particle. Ψ, is particularly important in quantum mechanics since a large number of properties can be derived from it.

Abbreviating $\frac{\partial^2}{\partial x^2}+\frac{\partial^2}{\partial y^2}+\frac{\partial^2}{\partial z^2}=\nabla^2$ and considering that in most cases we are interested in the time-independent potential, we can rewrite the Schrödinger equation as follows:

$$\left\{-\frac{\hbar^2}{2m}\nabla^2+V\right\}\Psi(r)=E\,\Psi(r) \text{ (Eq. 3.2)}$$

To solve the equation it is necessary to find the E values and the functions Ψ that satisfy the equality. This is known as an eigenvalue problem, where the energy is an eigenvalue of the operator, which is called the Hamilton operator. This equation, unfortunately, only has an exact solution for single-electron atoms; many-body systems must be solved by approximations and model assumptions.

The most widely recognized initial approach is the Hartree-Fock method.[4,5] In this method, the wave function is considered as a Slater determinant of orbitals for single particles.[6,7] Typically the orbitals of a hydrogenoid atom (an atom with any nuclear charge but only one electron). For many-body calculations, the initial wave functions are typically a linear combination of atomic orbitals. Once an initial wave function has been constructed, the average of the effect of all the other electrons is calculated and used to generate a potential. This gives an electron in a defined field, for which the Schrödinger equation can be solved. The procedure is repeated for each of the other electrons, until a complete cycle of the method is completed. Thereafter, iterations of the above procedure are performed until convergence. The Hartree-Fock method, however, by averaging the potential that each electron sees completely omits the correlation energy, which is not negligible. This is why Hartree-Fock is often the starting point for other ab initio methods, called post-Hartree-Fock methods, which introduce modifications to the method to try to include correlation effects.

On the other hand we have molecular mechanics. This is a method that allows the reduction of the complexity of the system by means of a series of assumptions. In short, the system is represented by a group of charged spheres (atoms with implicit electrons) joined by springs (chemical bonds). By this approach, the system no longer requires a quantum treatment of the electrons and can be modeled more cheaply, in terms of computational resources and time, by classical mechanics.[8] More details of this will be explained later.

Molecular mechanics and quantum mechanics are a good pool of methods for dealing with reactivity problems in biosystems. On the one hand, the computational efficiency of molecular mechanics methods can be combined to simulate longer times in equilibrium states in larger systems, allowing a more complete exploration of the configurational space. On the other hand, quantum mechanics by considering electrons explicitly allows the study of phenomena that depend on the electronic structure and its changes (reactivity). Quantum mechanics and molecular mechanics methods can in fact be combined in a single (hybrid quantum mechanics/molecular mechanics) model to complement their advantages. In this chapter there is an introductory description of the aforementioned.

## 3.2 ENERGY EVALUATION

### 3.2.1 SEMI-EMPIRICAL METHODS

The derivations of the Hartree-Fock equations for closed-shell systems are known as the Roothaan-Hall equations.[9,10] These equations are the cornerstone of *ab initio* methods. They are the most efficient way to firstly approach the practical solution of the Schrödinger equation for multielectronic systems. The Roothaan-Hall equations are:

$$\mathbf{FC}=\mathbf{SCE} \quad \text{(Eq. 3.3)}$$

where **S** is the overlap matrix, **C** is the coefficient matrix, **E** is the energy matrix and **F** is the Fock matrix with elements:

$$F_{\mu\nu}=H_{\mu\nu}^{\text{core}}+\sum_{\lambda=1}^{K}\sum_{\sigma=1}^{K}P_{\lambda\sigma}\left[(\mu\nu|\lambda\sigma)-\frac{1}{2}(\mu\lambda|\nu\sigma)\right] \quad \text{(Eq. 3.4)}$$

Where $P_{\lambda\sigma}$ corresponds to:

$$P_{\lambda\sigma}=2\sum_{i=1}^{N/2}c_{\lambda i}c_{\sigma i} \quad \text{(Eq. 3.5)}$$

And $H_{\mu\nu}^{\text{core}}$ is:

$$H_{\mu\nu}^{\text{core}}=\int d\nu_1\phi_\mu(1)\left[\frac{-1}{2}\nabla^2-\frac{\sum Z_A}{|r_1-RA|}\right]\phi_\nu(1) \quad \text{(Eq. 3.6)}$$

It is well known that the most expensive part of *ab initio* calculations is always the computation and manipulation of integrals. In an *ab initio* calculation, all elements of the Fock matrix are solved analytically. This, no matter where the electron pair is located, in the same atom, in bonded atoms or in far away separated atoms. It is evident, then, that a good way to reduce the computational cost is to make approximations or look for strategies to deal with the integrals in a more efficient way. Semi-empirical methods tackle this problem by explicitly solving only the integrals of the electrons in the valence electrons. Moreover, semi-empirical methods take the overlap matrix S as the identity matrix. By this approach the equation 3.3 is transformed into an eigenvalue problem ($\mathbf{FC}=\mathbf{CE}$) and the Fock matrix can be diagonalized by simply giving the coefficients and energies.[1]

The basis of many semi-empirical methods is the zero differential overlap (ZDO) approximation.[11,12] In this, the overlap between pairs of different orbitals ($\phi$) is set to zero directly. Thus, for all volume elements:

$$\phi_{\mu}\phi_{\nu}=0 \text{ (Eq. 3.7)}$$

This results in any overlap integral being:

$$S_{\mu\nu}=\delta_{\mu\nu} \text{ (Eq. 3.8)}$$

Direct implementation of this approach is unfeasible because it violates some theoretical requirements. Modifications such as complete neglect of differential overlap (CNDO),[13] neglect of diatomic differential overlap (NDDO)[14] or modified neglect of diatomic overlap (MNDO)[15,16] were proposed, implemented and successfully used at the time. The PM3[17,18] and PM6[19] methods are based on the latter two. In these methods, the ZDO approximation is applied only to those orbitals found in different atoms. Moreover, contrary to ZDO, these bielectronic integrals for orbitals that are on different atoms are parametrically approximated to values that depend only on the identity and distance of the nuclei involved. AM1,[20] PM3 and PM6 semi-empirical methods are the evolution of parameters and improvements to some terms of the NDDO-based methods.

The PM family of methods is characterized by being parameterized by automated standardized parameterization procedures, unlike the AM family where the parameters were assigned by chemical "intuition". PM6 in particular is a reparameterization with emphasis on biological systems, although it has shown good performance on systems in general.

### 3.2.2 DENSITY FUNCTIONAL THEORY

Density functional theory (DFT)[21,22] has revolutionized computational chemistry. It provides a tool to solve electronic structure and energy in a computationally more efficient way than *ab initio* aproaches.[23] The pivot of this theory is the description of the energy not as an eigenvalue problem of the wave function, as in ab initio methods, but as a functional of the electronic probability density $E\nu[\rho]$. DFT is conceptually justified by the Hohenberg-Kohn (HK) theorems,[21] which can be summarized as follows:

1. The electronic density contains enough information to find the energy and the wave function of a system. Consequently, there is a one-to-one correspondence between energy and density. The above is justified in:

   a) The density integration defines the number of electrons.
   b) The cusps in the density function define the position of the nucleus.
   c) The height of this cusp defines the nuclear charge.
   d) With these data it is possible to know the Hamiltonian of a system.
   e) Knowing the Hamiltonian one potentially knows the wave function and with this the energy of the system.

2. Analogous to the variational principle for quantum theory, any test electron density will yield a higher energy value than the actual one associated with the ground state.

The energy of a system can be calculated using the expression for the Kohn-Sham DFT energy in the framework of the Born-Oppenheimer approximation.[24] It is expressed as the sum of the kinetic and electrostatic potential energies described as functionals of the electron density ($\rho_o$).

$$E_0 = \langle T[\rho_o]\rangle + \langle V_{Ne}[\rho_0]\rangle + \langle V_{ee}[\rho_0]\rangle \text{ (Eq. 3.9)}$$

The first term of the sum is the kinetic energy. The next two are electrostatic potential terms of electrons with nuclei (N) and with other electrons (e). To solve this equation the strategy is to split the energy into a portion that can be solved exactly and a small portion that requires the so-called exchange-correlation functional. In this approach, a reference system is assumed which is thought as an electronically non-interacting state where the density is the same as in the real system in the ground state. Given the assumption of non-interacting electrons the integrals are easily solved and therefore the reference term is of exact solution. The rest of the energy will be treated as a perturbation to the reference system by means of the exchange-correlation functional.

On this basis, the first term of equation 3.9 can be rewritten as:

$$\langle T[\rho_0]\rangle = \langle T[\rho_0]\rangle_{per} + \langle T[\rho_0]\rangle_{ref} = \langle T[\rho_0]\rangle_{per} - \frac{1}{2}\sum_{i=1}^{2n} \langle \psi_1^{KS}(1)|\nabla^2|\psi_1^{KS}(1)\rangle \text{ (Eq. 3.10)}$$

where $\psi^{KS}$ are orbitals introduced to solve the kinetic energy, and to obtain the electron density by means of:

$$\rho_0 = \sum_i |\psi_i|^2 \text{ (Eq. 3.11)}$$

The last term of equation 3.9, the potential energy electron-electron, can be written as:

$$\langle V[\rho_0]\rangle = \langle V[\rho_0]\rangle_{per} + \langle V[\rho_0]\rangle_{ref} = \langle V[\rho_0]\rangle_{per} + \frac{1}{2}\iint \frac{\rho_0(r_1)\rho_0(r_2)}{r_{12}} dr_1 dr_2 \text{ (Eq. 3.12)}$$

Where each of them is written as the sum of the energy of the non-interacting reference system plus a perturbation term corresponding to the energy caused by the interaction. In both cases, kinetic and potential, we have exact solutions for the terms of reference. In the kinetic case it is solved using Kohn-Sham orbitals[22] (and function as Slater determinant) and the potential term as classical Coulombic interactions. The remaining perturbation terms contribute to the exchange-correlation functional.

$$E_{XC}=\langle T[\rho_0]\rangle_{per}+\langle V[\rho_0]\rangle_{per} \text{ (Eq. 3.13)}$$

There is no known exact expression for this term and it must therefore be approximated. There are several types of functionals with increasingly better approximations and hybrid approximations calculated by Hartree-Fock.

**Local density, local spin density and generalized gradient approximations.** Different approaches have been proposed to tackle the description of electron density. In the simplest of these, Local Density Approximation (LDA), the electron density is treated as a uniform gas cloud. This approximation was quickly replaced by Local Spin Density Approximation (LSDA),[25] where the contributions of the $\alpha$ spin electrons and the $\beta$ spin electrons are considered separately. These two approximations result in the same in closed-shell systems. The exchange term in the LSDA approximation can be derived from the Dirac formula as:

$$E_x^{LSDA}[\rho]=-2^{2/3}C_x\int\left(\rho_\alpha^{4/3}+\rho_\beta^{4/3}\right)dr \text{ (Eq. 3.14)}$$

where Cx is a numerical constant and $\rho_\alpha$ and $\rho_\beta$ corresponds to the electronic densities for the $\alpha$ and $\beta$ spin electrons, respectively.

On the other hand, the correlation energy does not have an analytical expression and will always be approximated by parametric interpolations. In general, the LDSA approximation underestimates the exchange energy by about 10% and overestimates the correlation energy by a factor of 2. Given this, its use in reactivity today is practically nule. A factor of 2 in the correlation energy can misestimate a bond strength by about 25 kcal/mol. Thus, the simplest useful approximation today corresponds to the functionals developed within the framework of the general gradient approximation (GGA).[26,27] In this approach, in order to consider the density as a non-uniform gas, the gradient of the density is included in addition to the density itself. The best known and best performing GGA functionals are the exchange functional developed by Becke (B88),[28] which is often combined with the correlation functional developed by Lee-Yang-Parr (LYP)[2930] and the exchange and correlation functionals developed by Perdew-Burke-Ernzerhof (PBE).[31]

**Gaussian/plane waves approach and pseudopotentials.** The computational power available today limits the ability to simulate molecular dynamics with high levels of theory such as DFT. Thus, in order to improve the performance, and thus to carry out mechano-quantum simulations efficiently, the Gaussian plane wave approximation (GPW)[32] was implemented in the framework of the DFT theory. In this, briefly, the electron density is described by two mathematical representations. The first one n(r), an atom-centered expansion of contracted Gaussian functions, typical for quantum calculations, and the second one ñ(r), auxiliary plane-wave bases as

$$\tilde{n}(r)=\frac{1}{\Omega}\sum_{G}\tilde{n}(G)e^{iG\cdot r} \quad \text{(Eq. 3.15)}$$

where Ω is the unit cell volume, and G are the reciprocal lattice vectors. The expansion coefficients are such that the two representations n(r) and ñ(r) are equal within the cell volume. This allows fast conversion between n(r), ñ(r) and ñ(G) using efficient mapping procedures and fast Fourier transforms (FFT).

In addition, since the wave function of the inner electrons is oscillating, especially near the nucleus, describing them by plane wave functions would require very large basis sets. Thus, the GWP is usually implemented by pseudopotentials (Figure 3.1) for modeling the inner electrons (in our case from Goedecker, Teter, and Hutter (GTH)).[33,34] As shown in Figure 3.1, the wave function fluctuations of the electronic core (radii smaller than $r_c$) are treated by adding a continuous potential that smoothes them. By reducing the representation of the atomic core to this kind of "ions" (nucleus plus internal electrons), we can simplify the basis sets used without losing the chemistry of the system.

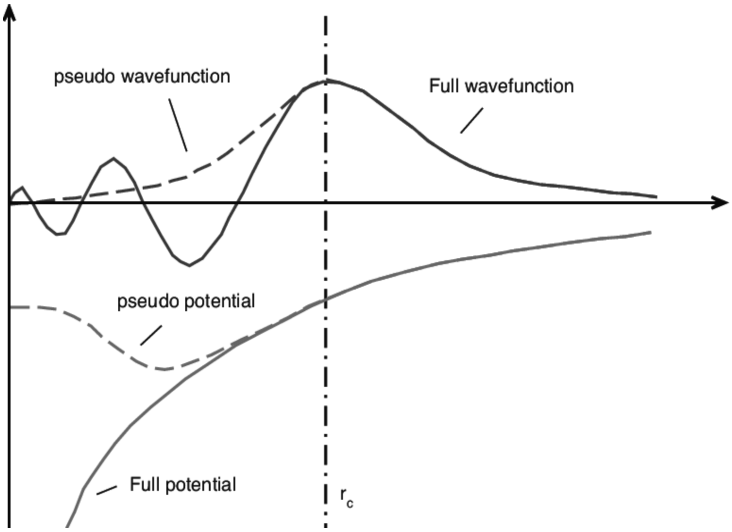


***Figure 3.1.*** *Schematic representation of a pseudopotential an its associated wavefuntion.* Taken *from* [1]

The Kohn-Sham DFT energy is defined within the GPW model by the expression

$$E(n)=E^{T}[n]+E^{V}[n]+E^{XC}[\tilde{n}]+E^{II}[\tilde{n}] \text{ (Eq. 3.16)}$$

where $E^T$ is the electronic kinetic energy, $E^V$ is the interaction between electrons and ionic cores, $E^{XC}$ is the correlation exchange energy, and $E^{II}$ corresponds to the interaction energy between ionic cores. As can be seen, the strength of this approximation is that each term of the equation is solved in the representation of the space in which it is most convenient to compute. The final wave function is computed by transforming between the spaces using fast Fourier transforms to the space where it is of interest. Usually, for analysis of electronic and physicochemical properties it is taken to the Gaussian wave space, since the shape of the orbitals has a representation closer to the chemical conception.

### 3.2.3 Empirical Force Field Models

In mechano-classical models, or empirical force fields (FF) models, the minimal unit of model construction is atoms as discrete particles. The omission of the electronic structure implies that all binding and charge information must be supplied somewhat implicitly as initial parameters and kept fixed throughout the calculation. Instead of being obtained cycle by cycle as in the mechano-quantum models. Additionally, omitting the solution of the electronic structure implies that it is not possible the description of any kind of phenomena that involves electrons such the bond breaking or forming processes. Moreover, nuclear motion is limited to follow only classical aspects. In other words, a calculation with force fields is reduced to obtaining energies for a certain arrangement of charged mass spheres (atoms) connected by springs (bonds), see Figure 3.2.[8]

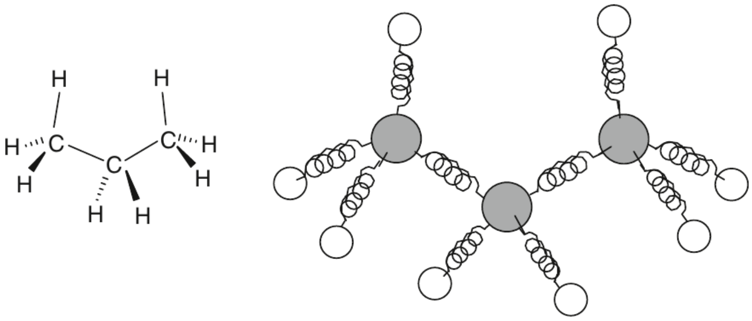


***Figure 3.2.*** *Propane molecule in a force-field-like representation. Bonds are represented by means of springs and atoms as charged spheres. Angles and dihedrals are computed as harmonic oscillators as a consequence of the mechano-classical description of bonds and atoms. Taken from* [8]

Thus, the energy calculated by means of a force field is given by the sum of terms described in equation 3.17. Each term describes the energy required to distort the arrangement of spheres bound together in some particular way. Or, alternatively, the electrostatic energy caused by the interaction of the charges:

$$E_{FF} = E_{bond} + E_{ang} + E_{tors} + E_{vdw} + E_{elec} + E_{cross} \text{ (Eq. 3.17)}$$

The first two terms, $E_{bond}$ and $E_{ang}$, correspond to the energies associated with distortions at equilibrium positions for bond distances and angles between bonding atoms. These are described by harmonic oscillator approximations as shown in Eq. 3.18 with parametric elastic constants ($k_{bond}$ and $k_{ang}$) and equilibrium values ($l_{eq}$ and $\theta_{eq}$).

$$E_{bond} = \sum_{bonds} k_{bond} (l - l_{eq})^2 \text{ and } E_{ang} = \sum_{angles} k_{ang} (\theta - \theta_{eq})^2 \text{ (Eq. 3.18)}$$

The element of the sum $E_{tors}$ is a term that describes the variations in energy due to rotations about dihedral and improper angles by combinations of cosines. In this way the symmetric nature of rotation about a dihedral can be captured.

$$E_{tors} = \sum_{r=1}^{dihedrals} k_r (1 + \cos(r\,\theta)) \text{ (Eq. 3.19)}$$

Where $k_r$ is the oscillation constant, r is an integer defining the periodicity of rotation and θ is the value of the dihedral or improper angle.

The next three elements of the energy are, first, that of van der Waals interactions, which is modeled by a 12-6 Lennard-Jones potential (Equation 3.20).

$$E_{vdw} = \sum_{cutoff} k_{vdw} \left[ \left( \frac{\sigma}{r} \right)^{12} - \left( \frac{\sigma}{r} \right)^{6} \right] \text{ (Eq. 3.20)}$$

In which the summation is performed on the atoms at less than a defined radius (cutoff) from which the van der Waals energies are neglected. This speeds up the calculation by avoiding the computation of terms for which the energy is practically negligible. $k_{vdw}$ is known as well depth and is associated with the energetic depth of the minimum at its optimum point with respect to the energy of the atoms separated at infinity and σ which is the collision diameter for each atom.

Second, the contribution of electrostatic interactions that is expressed by the Coulomb equation (Equation 3.21) for charged particles.

$$E_{elec} = \sum_{i=1}^{N} \sum_{j \neq 1}^{N-1} \frac{q_i q_j}{4 \pi \epsilon_0 r_{ij}} \quad \text{(Eq. 3.21)}$$

where $q_i$ and $q_j$ represent the charges of atoms i and j, respectively, separated by a distance $r_{ij}$.

And third, the cross term is an aggregate expression particular for force field (there is no one standard expression) to covers coupling between the previous mentioned fundamental, or diagonal, terms.

The constants, equilibrium values and loads mentioned in each of the above equations are stored as parameters in libraries called force fields. These libraries compile the data based on high-level theoretical calculations or experimental considerations. By this way, different force fields are created that will better or worse address the needs of the system to be modeled. For the study of large systems, such as proteins, there are FFs such as AMBER,[35] CHARMM,[36] GROMOS,[37] OPLS,[38] among others.

### 3.2.4 MULTISCALE MODELS

It is usual in computational biochemistry that modeling chemical processes requires the integration of quantum and classical potentials. Quantum mechanical models combined with molecular mechanics (QM/MM),[39–41] when applied to enzymatic systems, integrate the advantages of each method separately.

On the one hand, QM methods allow detailed analysis of electronic structure-dependent processes i.e. chemical reactions. This is implemented in the active site region of the enzyme, where the reactions of interest for the study of enzymatic catalysis or inhibition occur. On the other hand, MM methods can be implemented to systems composed of thousands of atoms, where computational costs would make the application of QM potentials difficult. Thus, the combination of the two methods allows the evaluation of chemical reactions, considering the effect of protein environment on the processes that take place in the active site.

In general terms, the energy of a system simulated by QM/MM models can be divided into three parts. A QM energy ($E_{QM}$), calculated by any quantum potential (as described above). A MM energy ($E_{MM}$), described by an empirical force field. Finally, and being the axis of this section, the interaction between the QM and MM regions ($E_{QM/MM}$).

Two major schemes have been proposed to address the total energy in QM/MM methods; a subtractive scheme and an additive scheme.

In the subtractive method, energy is obtained in three steps. First the MM energy is calculated for the complete system. Then the energy QM for the QM region is calculated and added to the previous one ($E_{MM}$). Up to this point, the energy of the QM system has been erroneously considered twice, once MM and once QM. To correct this, a third term is subtracted from the above, the MM energy of the QM region. Mathematically the above is expressed as follows:

$$E_{(QM/MM)} = E_{MM}(QM+MM) + E_{QM}(QM) - E_{MM}(QM) \quad \text{(Eq. 3.22)}$$

The subtractive scheme offers advantages such as the facility to include more layers in the model without complicating the calculation too much. An example is the integrated molecular orbital (MO) and MM (ONIOM)[42] method which allows the combination of an 'n' number of successive layers and is a popular choice nowadays

(QM/QM/MM models), implemented in the software Gaussian. However, the need for force field parameters for the QM region disadvantages these methods which present convergence problems modelling certain geometries.

In the additive schemes the total energy is calculated by the MM energy of the MM region plus the QM energy of the QM region plus the coupling energy of the QM-MM regions.

$$E_{(QM/MM)} = E_{MM}(MM) + E_{QM}(QM) + E_{(QM-MM)}(QM+MM) \quad \text{(Eq. 3.23)}$$

It is noteworthy that in this scheme only the MM region is computed by means of empirical force fields, thus avoiding the need for adequate parameters at all times for the QM region.

The term coupling can be described by three different levels. Each one more accurate than the previous one:[43]

1. *Mechanical embedding.* It considers only the bounded and steric effects of the MM region on the QM region.

2. *Electrostatic embedding.* In addition, the effect of the electrostatic field, generated by the charges of the MM region, on the QM region is considered.

3. *Polarizable embedding.* The description of the MM atoms is slightly modified to allow the two regions to polarize each other.

In the mechanical coupling between the QM and MM regions, the bond, angles and torsions terms involving atoms of the QM region will be considered by the corresponding energy terms in the force field (Figure 3.3 a), b) and c)). The non-bonding terms (electrostatic and Van der Waals, Figure 3.3 a), b) and c)) are also evaluated by MM means. It should always be kept in mind that the Lennard-Jones parameters are not updated during the calculation, so it is good practice to avoid the MM boundary being within 6 Å of the “center” of the reactive region. If the QM region is interrupted (a bond is truncated) it will typically be replaced by the addition of an atom (commonly H) that completes the valence, an squeme called linked atom (Figure 3.3 e)).[39,41]

In electrostatic embending, an extra term is added to the mono-electronic Hamiltonian. Basically, the electrons in the QM region will see the point charges of the MM atoms (Figure 3.3 f)) as:

$$h_i^{QM-MM} = h_i^{QM} - \sum_{J}^{M} \frac{e^2 Q_J}{4\pi\epsilon_0 |r_i - R_J|} \quad \text{(Eq. 3.24)}$$

where $h_i^{QM}$ corresponds to the original monoelectronic operator (kinetic and electron-nucleus), $r_i$ and $R_J$ correspond to the coordinates of electron i in the QM region and atom J, of M atoms, in the MM region. The term $Q_J$ allows the electrons in the QM region to "see" the MM atoms. This implies that $Q_J$ can take partial or even negative charge values.

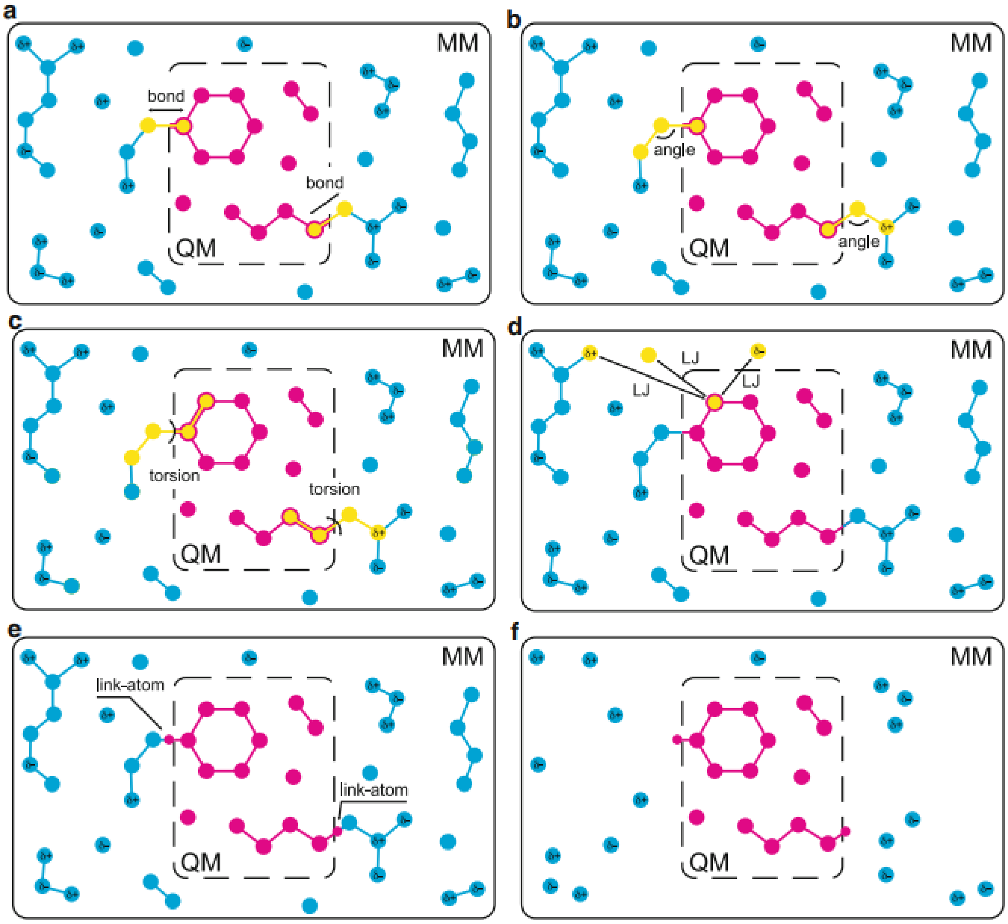


***Figure 3.3.*** *Schematic representation of coupling between the QM and MM subsystems in the additive QM/MM scheme. Panels represent a) bond, b) angles, c) torsion, d) Lennard-Jones terms evaluated by MM means just with the atoms inside a cutoff radii. e) Representation of the link atom scheme for the treatment of QM/MM frontiers and f) representation of the influence of the electrostatic field of the MM region (induced just by the charged atoms) in the QM sub-set.Taken from* [43]

Finally, although polarization embedding offers the most realistic coupling coupling there is not yet an effective implementation for biological systems. Therefore QM/MM studies with polarizable MM regions have so far been limited to non-biological systems.

### 3.3 POTENTIAL ENERGY SURFACES

Whenever one speaks of energy profiles, or reaction paths, it is common to refer to routes connecting critical points on the potential or free energy surfaces. An energy surface, either potential energy or free energy, is a landscape representation of energy as a function of coordinates of interest (usually one or two, for practicality). These coordinates, which can be internal or Cartesian, allow us to describe the evolution of a particular chemical process. The reactivity and other chemical behaviors of a system can be almost fully characterized just by the critical points (minima and maxima) of the associated energy surface. Therefore, being interested in minima or saddle points and their associated geometries is a typical problem in molecular modeling.

The problem of locating critical points consists of finding, for a potential energy function E that depends on one or more independent coordinates r1, r2, r3, ..., ri; the points where E has a minimum values. At a critical point, the gradient vector, **g** (elements defined by Equation 3.25), is zero. In addition, all components of the Hessian matrix (matrix of second derivatives), **H** (elements defined by Equation 3.25), are positive if it is a minimum, or all except one if it is a transition state.[8]

$$g_i = \frac{\partial E}{\partial r_i} \text{ and } H_{ij} = \frac{\partial^2 E}{\partial r_i \partial r_j} \text{ (Eq. 3.25)}$$

For systems with many atoms the surfaces can be extremely complex and therefore algorithms must be used to find these critical points. In general, these algorithms are called minimization algorithms (even if they refer to the location of a transition state) and can be of two types, derivative or non-derivative. Nowadays, derivative methods are the most commonly used, when both quantum and classical methodologies are used. These derivative methods, in turn, can be of first order or second order according to the order of the derivatives from which information is extracted for the refinement of the geometry.

Biomolecular systems are described by a large number of variables and therefore extremely complex state functions. This is why it is often necessary to use numerical methods to iterate from conformation to conformation until you get as close as you want to the critical point.

### 3.3.1 MINIMA AND SADDLE POINT LOCATION

**Minimization methods.** On the one hand, there are minimization algorithms that only allow us to go to lower points of the potential energy surface with respect to the initial structure. Therefore, they are used for the search of critical points of minimum value. The search for first order saddle points (transition states) are performed by other types of algorithms. A minimization calculation allows arriving at a single minimum on the surface. To search for several minima it is necessary to repeat the minimization from different initial structures.

The first-order derivative methods gradually modify the coordinates of the atoms to approach the minimum. For this they use the forces calculated with the first derivative of the energy function. In each iteration the initial geometry is the one obtained from the previous step. In the first of the iterative cycles the geometry is that provided by the user. The most commonly used first order derivative methods today are *steepest descent*[44] and *conjugate gradients*.[45]

In the case of steepest descent the displacements of the geometry are performed parallel to the net force. For a system of N particles, the components of the vector on which the displacement is performed is defined by

$$s_k = -\frac{g_k}{|g_k|} \text{ (Eq. 3.26)}$$

where $g_k$ corresponds to the gradient of the potential energy function. This is one of the most widely used methods due to the low computational cost. However, near minima where the surface is considerably flat, with small oscillations or with narrow valleys, this method leads to oscillatory behavior with multiple cycles without convergence. For such surfaces, the *conjugate gradients* method is more efficient.

*Conjugate gradients*, unlike *steepest descents*, performs the displacements along vectors of conjugate direction (thus the name of the method) and non-orthogonal. For this, the vector $v_k$, over which the system will move from a point defined by the coordinates of the vector $x_k$, is calculated from the current gradient and the direction of motion of the previous iteration ($V_{(k-1)}$) by

$$v_k = -g_k + \gamma_k v_{(k-1)} \quad \text{(Eq. 3.27)}$$

where $\gamma_k$ is a scalar defined by

$$\gamma_k = \frac{g_k \cdot g_k}{g_{k-1} \cdot g_{k-1}} \quad \text{(Eq. 3.28)}$$

Plainly, and given the definition of its terms, this method can only be implemented from the second iteration. Therefore, when the *conjugate gradients* method is invoked in a calculation the first cycle is carried out only along the gradient, i.e. a *steepest descents* cycle.

First-order derivative methods are ideal for large systems, such as those commonly treated in biochemistry. This is because a second order derivative matrix (Hessian matrix) represents a storage problem in the machine for complex systems. However, for small organic or inorganic systems, usually treated fully with quantum mechanics, second order methods are more frequent. In systems treated with a high level of quantum theory, the energy calculation is the most expensive part and therefore reducing at all costs the number of cycles is more important. In these cases, a second order method is then more optimal.

**Methods to locate saddle points.** The description of transition states for chemical processes must be done quantumly due to the electronic structure of these geometries. However, saddle point optimization algorithms are not yet efficient. The most commonly used take advantage of the fact that the gradient at these points is zero as well. Thus, optimization methods, adapted to move towards the upper parts of the surface, can converge to the critical point structure. The problem with these approximations is that they only work when the initial structure is correctly positioned in the quadratic region of the saddle point. Otherwise, it is quite likely to fall to one of the minima surrounding the saddle point. This is why second order algorithms are commonly used for the location of the saddle points.

One of the most common method is the *Newton-Raphson* method, a second order derivative method. In this, not only the first derivatives but also the second derivatives are used to locate the critical point. The second derivatives provide information about the curvature of the surface. If the initial structure is well selected, the Hessian matrix will have a predominant negative eigenvector. Along which the maximization will be carried out. The *Newton-Raphson* method starts from a Taylor expansion around a reference point $r_k$ of the surface. At this point, the function is defined by the expansion

$$E(r)=E(r_k)+(r-r_r)E'(r_k)+\frac{1}{2}(r-r_k)^2E''(r_k)+\ldots \text{ (Eq. 3.29)}$$

where $E'(r_k)$, $E''(r_k)$ are the first and second derivatives of $E(r_k)$ and the first derivative $E'(r)$ can by expressed by

$$E'(r)=rE'(r_r)+(r-r_r)E''(r_k) \text{ (Eq. 3.30)}$$

Then, if the function is pure quadratic, the second derivative is the same at any point, and thus

$$E''(r)=E''(r_k) \text{ (Eq. 3.31)}$$

And at the critical point ($x=x^*$) and

$$E'(r^*)=0 \text{ and } r^*=r_k-E'(r_k)E''^{-1}(r_k) \text{ (Eq. 3.32)}$$

where for a multidimensional function, $E''^{-1}(r_k)$ corresponds to the Hessian matrix, which must be inverted.

The computation of the Hessian matrix turns out to be the most expensive part along with the storage of it. Thus, the *Quasi-Newton* method,[46] the most widely used along with the *Broyden-Fletcher-Goldfarb-Shanno (BFGS)* method, approximates the Hessian matrix based on the computed gradients.

When the proposition of a correct initial structure to locate the saddle point by the optimization methods described above is not straightforward, algorithms such as *grid search*[47] or *quadratic synchronous transit approach*[48] allow to explore the surface. These are used either to locate the saddle point directly, by a systematic variation of the coordinates, or to try to relocate the structure in the quadratic region of the surface to then follow with a minimization method.

## 3.4 MOLECULAR DYNAMICS SIMULATIONS

Geometry optimization and exploration of a single PES provide valuable information of a reaction mechanism, but it is not enough to describe the reaction of a large system, such us biological molecules, in a representative manner. Energy surfaces of large molecules often possesses many minima connected by low energy barriers. Then, to relate simulations with the observable behavior of a system requires the exploration of all the possible conformations. Consequently, in chemistry, and especially with biological molecules, it is desirable to explore the potential energy surface extensively. Knowing the diversity of conformations that, at a certain temperature, are thermodynamically accessible to the system is crucial for the understanding of the phenomena. For this purpose, it is useful to collect a series of temporally related points (trajectories) propagating initial coordinates and velocities through a finite number of time steps.[1]

According to the Born-Oppenheimer approximation,[24] the atomic nuclei are heavier than electrons and they move much more slowly than their electrons. Then, it is possible to assume that the position of the electrons inmmediately adapts the change around the position of the nuclei. Moreover, nuclei can be considered as classical particles. Thus, performing molecular dynamics of a system is often reduced to solving the equation of Newton's second law, F=ma for the nuclei, which in differential form can be written as

$$F=-\frac{dV}{dr}=m\frac{d^2r}{dt^2} \text{ (Eq. 3.33)}$$

Where r is a vector of dimension $3N_{atoms}$ and V is the potential energy at these coordinates of the system.

Thus, given a set of initial coordinates, $r_i$, it is possible to know the coordinates ($r_{i+1}$) at a small time step later , $\Delta t$, using a Taylor expansion.

$$r_{i+1}=r_i+\frac{dr}{dt}(\Delta t)+\frac{1}{2}\frac{d^2r}{dt^2}(\Delta t)^2+\frac{1}{6}\frac{d^3r}{dt^3}(\Delta t)^3+\ldots \text{ (Eq. 3.34)}$$

And analogously, it is possible to know the coordinates of the previous step ($r_{i-1}$) in a small time step before using negative values of $\Delta t$

$$r_{i-1}=r_i-\frac{dr}{dt}(\Delta t)+\frac{1}{2}\frac{d^2r}{dt^2}(\Delta t)^2-\frac{1}{6}\frac{d^3r}{dt^3}(\Delta t)^3+\ldots \text{ (Eq. 3.35)}$$

By summing the above equations and truncating in the third term of the expansion, we obtain the main equations of Verlet's algorithm[49] for the propagation of molecular dynamics

$$r_{i+1}=(2r_i-r_{i-1})+a_i(\Delta t)^2 \text{ and } a_i=\frac{F_i}{m_i}=-\frac{1}{m_i}\frac{dV}{dr_i} \text{ (Eq. 3.36)}$$

Thus, using the positions $r_i$, the acceleration $a_i$ and the positions of the previous step $r_{i-1}$ to calculate the new positions $r_{i+1}$.

In the Verlet algorithm the velocities do not appear explicitly. In order to be able to make the implementation of a thermostat efficient, one must have explicit knowledge of the velocities. Therefore, a procedure analogous to that of the positions is carried out to arrive at the equations of the Verlet algorithm for velocity propagation,

$$v_{i+1}=v_i+\frac{1}{2}(a_i+a_{i+1})\Delta t \text{ (Eq. 3.37)}$$

having then complete the base system of equations by means of which the trajectory calculation is carried out.

The advantage of considering the previous and following states is that, in general, these dynamics have a tendency to show better energy distribution and conservation over very long simulation times, contrary to what happens in algorithms such as Runge-Kutta.[50]

The time steps used in the above equations are especially important. Since not considering how the forces vary during displacement, erroneous events can be simulated. These steps cannot exceed the fastest vibration time in the system. In the case of biomolecules, these are the vibrations of the bonds of heavy atoms with hydrogens, which have oscillation times of about one femtosecond. This implies that simulations must be carried out by collecting points every 0.5-2 femtoseconds of time. This limits, in view of the computational cost, affordable simulations to be between pico and a few microseconds.

**Temperature and pressure control.** It is common in molecular modeling to wish to carry out the simulation with certain constant conditions. Although many thermodynamic properties can be transformed between ensembles, this only applies to systems of infinite size ('the thermodynamic limit'). Therefore, it is usually more convenient to directly compute on the set where the analysis is to be performed. In chemistry, the most desirable ensembles are NPT and NVT due to their relation with the Gibbs and Helmholtz free energies, respectively, which are determining properties in the chemical behavior of the systems.[1]

To begin with, temperature is a more intuitive property to control in molecular dynamics. If we start from its thermodynamic definition, temperature is a measure of the translational kinetic energy of particles. So, directly, we can think of controlling the temperature by regulating the velocity of the particles that compose the system. The expression that analytically relates the kinetic energy to the temperature is

$$\langle \mathrm{K} \rangle = \frac{3}{2} \mathrm{N} k_{\mathrm{B}} \mathrm{T} \text{ (Eq. 3.38)}$$

where N is the number of particles, $k_B$ is the Boltzmann constant and T is the temperature of the system.

Then, the temperature T(t) at a time t can be rescaled by multiplying the velocities (v) of the N particles in the system by a factor λ, the resulting change is expressed by,

$$\Delta T=\frac{1}{2}\sum_{i=1}^{N}\frac{2}{3}\frac{m_i(\lambda v_i)^2}{Nk_B}-\frac{1}{2}\sum_{i=1}^{N}\frac{2}{3}\frac{m_i v_i^2}{Nk_B}$$

$$\Delta T=(\lambda^2-1)T(t)\ \text{with}\ \lambda=\sqrt{\frac{T_{new}}{T}(t)}\ \text{(Eq. 3.39)}$$

Where ΔT is the change derived in the temperature from rescaling to the reference, $m_i$ the mass of particle i with velocity $v_i$.

With this expression of λ we have a factor that quickly allows us to rescale the velocities knowing the current temperature of the system and the one at which we want to maintain it.

Unfortunately, although this method is simple, it is not very efficient. Since it rescales all velocities identically, it prolongs any temperature differences in the system. This is not convenient for our purposes since it is common throughout the research protocols to perform minimizations, which then lead to low forces and low initial velocities. An efficient thermostat, then, must also be able to redistribute kinetic energy while regulating velocities. Two methods are known today capable of this, the *stochastic collisions* method[51] initially proposed by Andersen and the *extended system* method proposed by Nosé/Hoover.[52]

In the *stochastic collision* method (the one we used) a particle is randomly selected every few intervals and its velocity is randomly reassigned by choosing a value within a Maxwell-Boltzmann distribution. The name of the method is due to the fact that we couple the system with a thermal bath that randomly emits thermal particles that collide with the atoms. By calculating the change in energy due to each collision, Andersen deduced that the average rate (v) at which each particle must undergo a collision is given by

$$v=\frac{2a\kappa}{3k_B N_d^{1/3} N^{1/3}} \text{ (Eq. 3.40)}$$

where a is a dimensionless constant, κ is the thermal conductivity and $N_d$ is the number density of the particles. If the thermal conductivity is not available, it can be predicted by the intermolecular collision frequency ($v_c$)

$$v=\frac{v_c}{N^{2/3}} \text{ (Eq. 3.41)}$$

It should always be taken into consideration that very high collision frequency values will prevent the system from undergoing the normal kinetic energy fluctuations and therefore reduce the phase space sampling. If, on the other hand, the collision frequency is too low, there is a risk of not sampling at a constant temperature.

Pressure proves to be a much more fluctuating variable in molecular dynamics. This is not unexpected as the pressure is obtained by means of the virial, which in turn is obtained as the product between the positions and the derivative of the potential energy function. The product, ($r_{ij} dV(r_{ij})/dr_{ij}$), fluctuates faster than other observables as the internal energy or the temperature. But, although its fluctuations are higher, its average is perfectly adjustable by controlling the volume (the position of the particles) of the system with a certain frequency. The magnitude of the change in volume will be given by the isothermal compressibility constant κ:

$$\kappa=\frac{-1}{V}\left(\frac{\partial V}{\partial P}\right)_T \text{ (Eq. 3.42)}$$

where V is the volume, P is the pressure and T is the temperature.

Also, isothermal compressibility is related to volume by means of:

$$\kappa = \frac{1}{k_B T} \frac{\langle V^2 \rangle - \langle V \rangle^2}{\langle V^2 \rangle} \text{ (Eq. 3.43)}$$

Analogous to scaling the temperature, we can derive a factor λ by which we rescale the atomic coordinates, and with them the box size of the system (if the system has been introduced in a box of solvent molecules), so that:

$$r_i^{new} = \lambda r_i \text{ (Eq. 3.44)}$$

where λ includes the isothermal compressibility information of the system.

## 3.5 FREE ENERGY CALCULATIONS

The free energy, in physical chemistry, is considered one of the most important quantities. In our case, free energy values allow us a direct validation of models by comparing activation energies with kinetic constants as was described in the introduction. Unfortunately, as well as important, it is also one of the most complicated to be estimated computationally for large flexible systems (such as biomolecules) due to the multiple configurations separated by low energy barriers. In addition, methods such as conventional molecular dynamics fail to correctly sample the higher energy states and therefore do not allow an accurate estimation of the free energy. this factor is even more problematic in reactivity since transition states are themselves high energy states unavailable to molecular dynamics. Therefore, it is necessary to employ variations in molecular dynamics in order to adequately access and sample those regions of higher energy that are of utmost importance. In this section some of the most recognized methods that are used to evaluate the free energy differences associated with different types of chemical transitions will be presented.

### 3.5.1 UMBRELLA SAMPLING

*Umbrella sampling*[53–55] is one of the most widely used methods to attack the problem of sampling along an interest transition. It is based on the addition of an extra potential that "traps" the system at different points of the transition, allowing it to sample zones higher in energy along the s(r) coordinate that defines the process (see Figure 3.4).

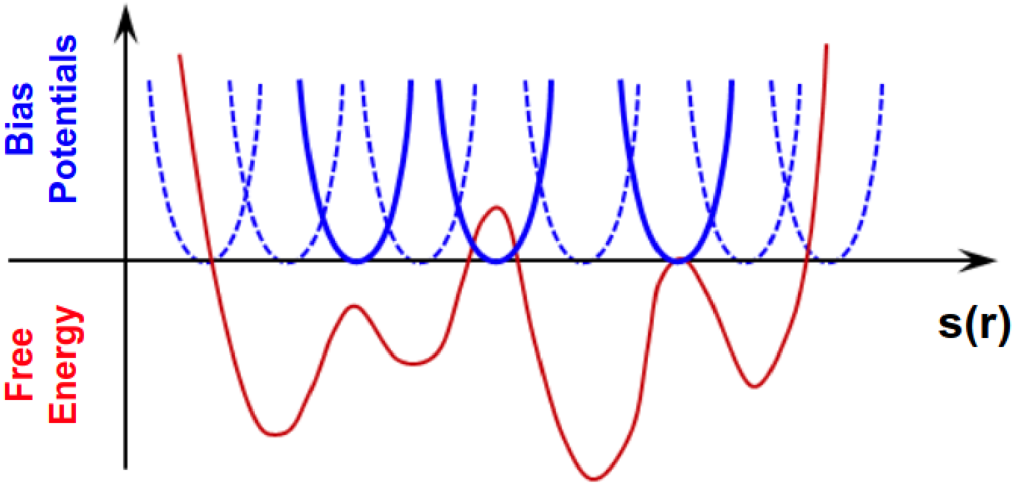


***Figure 3.4.*** *Schematic representation of the umbrella sampling approach. The blue lines represent the bias potentials added to the system at different s(r) values (windows) along the coordinate that space.*

The modified potential of the system can be written as:

$$V'(r)=V(r)+W(s(r)) \quad \text{(Eq. 3.45)}$$

where the bias potential W is expressed as an harmonic potential by:

$$W(s(r))=k_W(s-s_0)^2 \quad \text{(Eq. 3.46)}$$

where the conformations with coordinates r around the central configuration of coordinates $r_0$, where the system is "trapped", will be sampled. In this way, it is possible to carry out more complete sampling of the surfaces, which allows, by methods as *umbrella integration*[56] or *weighted histogram analysis method (WHAM)*,[57] to calculate properties such as the free energy associated with the chemical process.

### 3.5.2 METADYNAMICS CALCULATIONS

Metadynamics[58,59] is an enhanced sampling method as is *umbrella sampling.* In this method, a history-dependent bias potential is constructed during the molecular dynamics as a sum of repulsive Gaussians along the space of a collective variable (CV), s(r), (see Figure 3.5). These Gaussians are centered on the average value of the points explored in the CV up to the time of their addition. These have assigned widths ($\sigma$) and heights ($w_G$) that will be the same for all Gaussians added every $\tau_G$ time along the simulation progresses. Thus, the potential built up to a simulation time t is:

$$V(s(r),t)=\sum_{i=1}^{t/\tau_G} w_G \exp\left[-\sum_{\alpha=1}^{N_{CV}} \frac{s_\alpha - s_\alpha(i\tau_G)^2}{2\sigma_\alpha^2}\right] \text{(Eq. 3.47)}$$

for $\alpha$ collective variables biased in the simulation.

Note that the width must be set for each of the CVs ($\sigma_\alpha$). These widths determine the fit of each CV, i.e., they can be used to establish what distance in CV space should be considered non-visitable.

From Figure 3.5 it can be deduced that this method not only discourages exploration of states already visited in the CV space, but also provides an immediate estimate of the underlying free energy surface F(s) (black line in Figure 3.5). Since Gaussians are most likely to be added at points where the total effective free energy F(s)+V(s) (dashed gray line in Figure 3.5) is lower, the effect tends to flatten the F(s)+V(s) function. After a suitable "filling" time, the bias will begin to grow parallel to itself, and one can expect to directly estimate F(s) as the negative of V(s) (solid gray line in Figure 3.5).

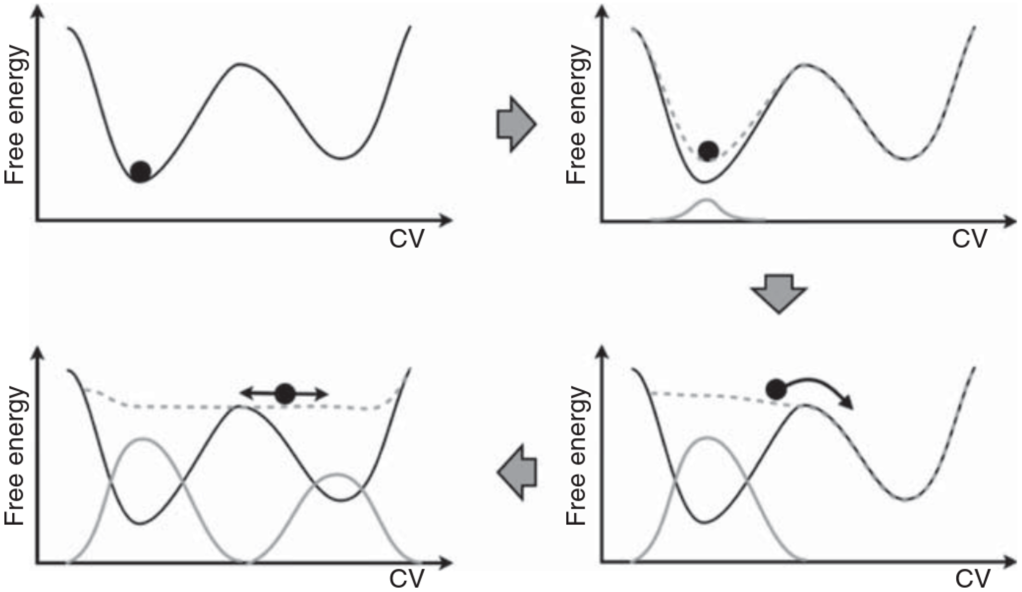


***Figure 3.5.*** *Schematic representation of the metadynamics approach progress. The solid gray line represents the added Gaussian bias potential. The dashed gray line shows the elevation of the free energy landscape caused by the addition of Gaussian bias potentials. The sum of the deposited bias (solid gray profile) provides a first rough negative estimate of the free energy profile. Taken from* [60]

Since metadynamics is a flat histogram method, it tries to sample the entire CV space. As a result, the simulated system can be pushed into states with a free energy that is not physically high and could skew the simulation toward thermodynamically irrelevant configurations. The addition of extra restrain potentials are often used to prevent this phenomenon.

### 3.5.3 THERMODYNAMIC INTEGRATION

One approach to obtain the free energy change between a initial state and final state is to use a coupling parameter or a descriptor λ.[61] This parameter will vary from 0 to 1 as the transition takes place. Thus, λ will have a value of 0 for the initial state and a value of 1 for the final state. Then, if the free energy, A, is a function of the parameter one can write that:

$$\Delta A = \int_0^1 \frac{\partial A(\lambda)}{\partial \lambda} d\lambda \text{ (Eq. 3.48)}$$

introducing the statistical thermodinamic definition of A

$$A(\lambda) = -k_B T \ln Q(\lambda) \text{ (Eq. 3.49)}$$

and the partition function for the cannonical ensamble ($Q_{NVT}$)

$$Q_{NVT} = \frac{1}{N!}\frac{1}{h^{3N}} \int\int dp^N dr^N \exp\left[\frac{-H(p^N, r^N)}{k_B T}\right] \text{ (Eq. 3.50)}$$

wher h is the Plank's constant and H is the Hamiltonian, function of the momentum p and the coordinates r of the N particles.

It is possible to deduce

$$\Delta A = \int_{\lambda=0}^{\lambda=1} \left\langle \frac{\partial H(p^N, r^N, \lambda)}{\partial \lambda} \right\rangle_\lambda d\lambda \text{ (Eq. 3.51)}$$

This means that to calculate the free energy difference between two coupled states by means of the λ parameter we only have to calculate this integral (Eq. 3.51). In practical terms this is done by performing multiple simulations at different discrete values of lambda between 0 and 1. Subsequently, in each simulation the average

$$\left\langle \frac{\partial H(p^N, r^N, \lambda)}{\partial \lambda} \right\rangle_\lambda \text{ (Eq. 3.52)}$$

is computed. The total free energy difference ΔA is then obtained as the sum of the area under the graph of values obtained from Equation 3.52 versus λ.

A particular case of thermodynamic integration is the alchemical transformations.[62] In this methodology, the coupling parameter $\lambda$ is used to calculate the energy differences associated with going from one chemical species to another. For this case, the coupling parameter scales the charges and the masses of the disappearing and appearing atoms that will be transmuted (see Figure 3.6).

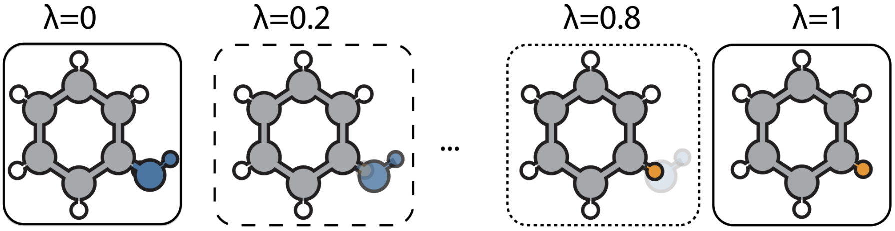


***Figure 3.6.*** *Representation of an alchemical transformation in which alchemical intermediates are created by making the potential energy dependent on an additional variable* λ *that interpolates between the chemical starting and end points. In this case the Figure shows a transformation from phenol to benzene. Adapted from* [63]

### 3.5.4 MM/PBSA APPROACH AND INTERACTION ENTROPY

In drug design it is very common to search for compounds that bind to a receptor, usually a protein. The binding process can be described by the chemical equation:

$$L + P \rightarrow PL \quad \text{(Eq. 3.53)}$$

where L is the ligand, P is the protein and PL is the ligand:protein complex.

The potency of a drug depends largely on the strength with which it binds to its target receptor. Therefore, a magnitude that allows quantifying the strength of this process is crucial for medicinal chemistry. That magnitude is the binding free energy, which based on the Equation 3.53 can be written as the difference of the free energy of the complex and the free energies of the ligand and the protein separated.

$$\Delta G_{bind} = \langle G_{PL} \rangle - \langle G_P \rangle - \langle G_L \rangle \quad \text{(Eq. 3.54)}$$

where the energy of each term can be estimated by the expression

$$G = E_{bnd} + E_{el} + E_{vdW} + G_{pol} + G_{np} - TS \text{ (Eq. 3.55)}$$

Where the first three terms correspond to I) the classical bonding terms ($E_{bond}$, $E_{ang}$ and $E_{tor}$), ii)electrostatic and iii) van der Waals MM values. The next two correspond to the polar and non-polar contributions to the solvation free energy. Finally, the last term (TS) corresponds to the entropic contribution in each state, which must be calculated by means of normal modes. Practically, the calculation of these free energies is approached by making some approximations. Instead of calculating the following terms

$$\Delta G_{bind} = \langle G_{PL} \rangle_{PL} - \langle G_P \rangle_P - \langle G_L \rangle_L \text{ (Eq. 3.56)}$$

where the subscripts indicate from which simulated system the term is obtained, it is usual to perform a single simulation of the complex protein:ligand and approximate to

$$\Delta G_{bind} = \langle G_{PL} - G_P - G_L \rangle_{PL} \text{ (Eq. 3.57)}$$

This not only reduces the number of simulations from three to one, but also cancels the $E_{bnd}$ term from the Equation 3.55. The next two terms are solved using implicid solvent models. The term $E_{pol}$ is obtained by solving the Poisson-Boltzmann equation for the description of electrostatic interactions of ions in solution or the generalized Born model for MM/GBSA.[64] The term $E_{np}$ is estimated from a linear relation to the solvent accessible surface area (SASA).[65] Finally, the entropic term is often the most costly, and usually many simply ignore it, actually reporting binding enthalpies $\Delta H_{bind}$.

**Interaction entropy.** An approximation that allows estimating part of the entropic contributions is called the interaction entropy approximation.[66] This allows estimating the -TS term of the equation 3.55 without having to solve the normal modes. For this approximation, we can reformulate the binding energy a bit and consider it as,

$$\Delta G_{bind} = \Delta G_{gas} + \Delta G_{sol} \text{ (Eq. 3.58)}$$

where the $\Delta G_{sol}$ term gathers the differences of the $\Delta G_{pol}$ and $\Delta G_{np}$ terms and therefore the $\Delta G_{gas}$ term corresponds to the MM and entropic terms.

$$\Delta G_{gas} = \langle E_{PL} \rangle - T \Delta S \text{ (Eq. 3.59)}$$

Now, if we start from the statistical definition of free energy, using the partition functions, we can write the energy $\Delta G_{gas}$ as:

$$\Delta G_{gas} = -k_B T \ln \frac{\int dq_w dq_p dq_l e^{-\beta(E_p + E_L + E_{PL} + E_W + E_{PW} + E_{LW})}}{\int dq_p dq_l dq_w e^{-\beta(E_P + E_L + E_W + E_{PW} + E_{LW})}}$$

$$\Delta G_{gas} = -k_B T \ln \left[ \frac{1}{\langle e^{\beta E_{PL}} \rangle} \right] = KT \ln \langle e^{\beta E_{PL}} \rangle$$

$$\Delta G_{gas} = k_B T \ln \left[ e^{\beta \langle E_{PL} \rangle} \langle e^{\beta (E_{PL} - \langle E_{PL} \rangle)} \rangle \right]$$

$$\Delta G_{gas} = \langle E_{PL} \rangle + k_B T \langle e^{\beta \Delta E_{PL}} \rangle \text{ (Eq. 3.60)}$$

where the classical water energy ($E_W$), protein:water interaction energy ($E_{PW}$), and ligand:water interaction energy ($E_{PW}$) are included. $\beta$ equals $1/k_B T$.

Which finally when compared with the Equation 3.59 results in the definition of the interaction entropy

$$-T\Delta S = k_B T\left\langle e^{\beta \Delta E_{PL}}\right\rangle \text{ (Eq. 3.61)}$$

Which can be easily obtained from the same simulation of the protein:ligand complex.

## 3.6 REFERENCES

(1) Leach, A. R. Molecular Modelling : Principles and Applications Second Edition. *Pearson Educ. EMA* **2001**.

(2) Schrödinger, E. Quantisierung Als Eigenwertproblem. *Ann. Phys.* **1926**, *384* (4), 361–376. https://doi.org/10.1002/andp.19263840404.

(3) Schrödinger, E. An Undulatory Theory of the Mechanics of Atoms and Molecules. *Phys. Rev.* **1926**, *28* (6), 1049–1070.

(4) Hartree, D. R. The Wave Mechanics of an Atom with a Non-Coulomb Central Field. Part I. Theory and Methods. *Math. Proc. Cambridge Philos. Soc.* **1928**, *24* (1), 89–110. https://doi.org/10.1017/S0305004100011919.

(5) Fock, V. Näherungsmethode Zur Lösung Des Quantenmechanischen Mehrkörperproblems. *Zeitschrift für Phys.* **1930**, *61* (1), 126–148. https://doi.org/10.1007/BF01340294.

(6) Slater, J. C. Note on Hartree's Method. *Phys. Rev.* **1930**, *35* (2), 210–211. https://doi.org/10.1103/PhysRev.35.210.2.

(7) Slater, J. C. A Simplification of the Hartree-Fock Method. *Phys. Rev.* **1951**, *81* (3), 385–390. https://doi.org/10.1103/PhysRev.81.385.

(8) Lewars E. G. *Computational Chemistry: Introduction to the Theory and Applications of Molecular and Quantum Mechanics*, Second Edi.; 2011. https://doi.org/10.1007/978-90-481-3862-3.

(9) Roothaan, C. C. J. New Developments in Molecular Orbital Theory. *Rev. Mod. Phys.* **1951**, *23* (2), 69–89. https://doi.org/10.1103/RevModPhys.23.69.

(10) Hall, G. G. The Molecular Orbital Theory of Chemical Valency VIII. A Method of Calculating Ionization Potentials. *Proc. R. Soc. London. Ser. A. Math. Phys. Sci.* **1951**, *205* (1083), 541–552. https://doi.org/10.1098/rspa.1951.0048.

(11) Pariser, R.; Parr, R. G. A Semi-Empirical Theory of the Electronic Spectra and Electronic Structure of Complex Unsaturated Molecules. II. *J. Chem. Phys.* **1953**, *21* (5), 767–776. https://doi.org/10.1063/1.1699030.

(12) Pople, J. A. Electron Interaction in Unsaturated Hydrocarbons. *Trans. Faraday Soc.* **1953**, *49*, 1375. https://doi.org/10.1039/tf9534901375.

(13) Pople, J. A.; Segal, G. A. Approximate Self-Consistent Molecular Orbital Theory. III. CNDO Results for AB 2 and AB 3 Systems. *J. Chem. Phys.* **1966**, *44* (9), 3289–3296. https://doi.org/10.1063/1.1727227.

(14) Pople, J. A.; Santry, D. P.; Segal, G. A. Approximate Self-Consistent Molecular Orbital Theory. I. Invariant Procedures. *J. Chem. Phys.* **1965**, *43* (10), S129–S135. https://doi.org/10.1063/1.1701475.

(15) Dewar, M. J. S.; Thiel, W. Ground States of Molecules. 38. The MNDO Method. Approximations and Parameters. *J. Am. Chem. Soc.* **1977**, *99* (15), 4899–4907. https://doi.org/10.1021/ja00457a004.

(16) Dewar, M. J. S.; Thiel, W. Ground States of Molecules. 39. MNDO Results for Molecules Containing Hydrogen, Carbon, Nitrogen, and Oxygen. *J. Am. Chem. Soc.* **1977**, *99* (15), 4907–4917. https://doi.org/10.1021/ja00457a005.

(17) Stewart, J. J. P. Optimization of Parameters for Semiempirical Methods I. Method. *J. Comput. Chem.* **1989**, *10* (2), 209–220. https://doi.org/10.1002/jcc.540100208.

(18) Stewart, J. J. P. Optimization of Parameters for Semiempirical Methods II. Applications. *J. Comput. Chem.* **1989**, *10* (2), 221–264. https://doi.org/10.1002/jcc.540100209.

(19) Stewart, J. J. P. Optimization of Parameters for Semiempirical Methods V: Modification of NDDO Approximations and Application to 70 Elements. *J. Mol. Model.* **2007**, *13* (12), 1173–1213. https://doi.org/10.1007/S00894-007-0233-4/FIGURES/10.

(20) Dewar, M. J. S.; Zoebisch, E. G.; Healy, E. F.; Stewart, J. J. P. Development and Use of Quantum Mechanical Molecular Models. 76. AM1: A New General Purpose Quantum Mechanical Molecular Model. *J. Am. Chem. Soc.* **1985**, *107* (13), 3902–3909. https://doi.org/10.1021/ja00299a024.

(21) Hohenberg, P.; Kohn, W. Density Functional Theory. *Phys. Rev.* **1964**, *136* (3B), B864–B871. https://doi.org/10.1103/PhysRev.136.B864.

(22) Kohn, W.; Sham, L. J. Self-Consistent Equations Including Exchange and Correlation Effects. *Phys. Rev.* **1965**, *140* (4A). https://doi.org/10.1103/PhysRev.140.A1133.

(23) Jensen, F. *Introduction to Computational Chemistry*, Third Edit.; 2017.

(24) Born, M.; Oppenheimer, R. Zur Quantentheorie Der Molekeln. *Ann. Phys.* **1927**, *389* (20), 457–484. https://doi.org/10.1002/andp.19273892002.

(25) Perdew, J. P.; Zunger, A. Self-Interaction Correction to Density-Functional Approximations for Many-Electron Systems. *Phys. Rev. B* **1981**, *23* (10), 5048–5079. https://doi.org/10.1103/PhysRevB.23.5048.

(26) Becke, A. D. Density-functional Thermochemistry. II. The Effect of the Perdew–Wang Generalized-gradient Correlation Correction. *J. Chem. Phys.* **1992**, *97* (12), 9173–9177. https://doi.org/10.1063/1.463343.

(27) Perdew, J. P.; Chevary, J. A.; Vosko, S. H.; Jackson, K. A.; Pederson, M. R.; Singh, D. J.; Fiolhais, C. Atoms, Molecules, Solids, and Surfaces: Applications of the Generalized Gradient Approximation for Exchange and Correlation. *Phys. Rev. B* **1992**, *46* (11), 6671–6687. https://doi.org/10.1103/PhysRevB.46.6671.

(28) Becke, A. D. Density-Functional Exchange-Energy Approximation with Correct Asymptotic Behavior. *Phys. Rev. A* **1988**, *38* (6), 3098–3100. https://doi.org/10.1103/PhysRevA.38.3098.

(29) 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* **1988**, *37* (2), 785. https://doi.org/10.1103/PhysRevB.37.785.

(30) Miehlich, B.; Savin, A.; Stoll, H.; Preuss, H. Results Obtained with the Correlation Energy Density Functionals of Becke and Lee, Yang and Parr. *Chem. Phys. Lett.* **1989**, *157* (3), 200–206. https://doi.org/10.1016/0009-2614(89)87234-3.

(31) Perdew, J. P.; Burke, K.; Ernzerhof, M. Generalized Gradient Approximation Made Simple. *Phys. Rev. Lett.* **1996**, *77* (18), 3865–3868. https://doi.org/10.1103/PhysRevLett.77.3865.

(32) Lippert, G.; Hutter, J.; Parrinello, M. A Hybrid Gaussian and Plane Wave Density Functional Scheme. *Mol. Phys.* **1997**, *92* (3), 477–487. https://doi.org/10.1080/002689797709482119.

(33) Krack, M. Pseudopotentials for H to Kr Optimized for Gradient-Corrected Exchange-Correlation Functionals. *Theor. Chem. Acc.* **2005**, *114* (1–3), 145–152. https://doi.org/10.1007/s00214-005-0655-y.

(34) Goedecker, S.; Teter, M. Separable Dual-Space Gaussian Pseudopotentials. *Phys. Rev. B - Condens. Matter Mater. Phys.* **1996**, *54* (3), 1703–1710. https://doi.org/10.1103/PhysRevB.54.1703.

(35) Salomon-Ferrer, R.; Case, D. A.; Walker, R. C. An Overview of the Amber Biomolecular Simulation Package. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2013**, *3* (2), 198–210. https://doi.org/10.1002/wcms.1121.

(36) Best, R. B.; Zhu, X.; Shim, J.; Lopes, P. E. M.; Mittal, J.; Feig, M.; MacKerell, A. D. Optimization of the Additive CHARMM All-Atom Protein Force Field Targeting Improved Sampling of the Backbone ϕ, ψ and Side-Chain χ 1 and χ 2 Dihedral Angles. *J. Chem. Theory Comput.* **2012**, *8* (9), 3257–3273. https://doi.org/10.1021/ct300400x.

(37) Oostenbrink, C.; Villa, A.; Mark, A. E.; Van Gunsteren, W. F. A Biomolecular Force Field Based on the Free Enthalpy of Hydration and Solvation: The GROMOS Force-Field Parameter Sets 53A5 and 53A6. *J. Comput. Chem.* **2004**, *25* (13), 1656–1676. https://doi.org/10.1002/jcc.20090.

(38) Jorgensen, W. L.; Tirado-Rives, J. The OPLS [Optimized Potentials for Liquid Simulations] Potential Functions for Proteins, Energy Minimizations for Crystals of Cyclic Peptides and Crambin. *J. Am. Chem. Soc.* **1988**, *110* (6), 1657–1666. https://doi.org/10.1021/ja00214a001.

(39) Field, M. J.; Bash, P. A.; Karplus, M. A Combined Quantum Mechanical and Molecular Mechanical Potential for Molecular Dynamics Simulations. *J. Comput. Chem.* **1990**, *11* (6), 700–733. https://doi.org/10.1002/jcc.540110605.

(40) Warshel, A.; Levitt, M. Theoretical Studies of Enzymic Reactions: Dielectric, Electrostatic and Steric Stabilization of the Carbonium Ion in the Reaction of Lysozyme. *J. Mol. Biol.* **1976**, *103* (2), 227–249. https://doi.org/10.1016/0022-2836(76)90311-9.

(41) Singh, U. C.; Kollman, P. A. A Combined Ab Initio Quantum Mechanical and Molecular Mechanical Method for Carrying out Simulations on Complex Molecular Systems: Applications to the CH3Cl + Cl? Exchange Reaction and Gas Phase Protonation of Polyethers. *J. Comput. Chem.* **1986**, *7* (6), 718–730. https://doi.org/10.1002/jcc.540070604.

(42) Dapprich, S.; Komáromi, I.; Byun, K. S.; Morokuma, K.; Frisch, M. J. A New ONIOM Implementation in Gaussian98. Part I. The Calculation of Energies, Gradients, Vibrational Frequencies and Electric Field Derivatives. *J. Mol. Struct. THEOCHEM* **1999**, *461–462*, 1–21. https://doi.org/10.1016/S0166-1280(98)00475-8.

(43) Groenhof, G. Introduction to QM/MM Simulations; 2013; pp 43–66. https://doi.org/10.1007/978-1-62703-017-5_3.

(44) Haug, E. J.; Arora, J. S.; Matsui, K. A Steepest-Descent Method for Optimization of Mechanical Systems. *J. Optim. Theory Appl.* **1976**, *19* (3), 401–424. https://doi.org/10.1007/BF00941484.

(45) Hestenes, M. R.; Stiefel, E. Methods of Conjugate Gradients for Solving Linear Systems. *J. Res. Natl. Bur. Stand. (1934).* **1952**, *49* (6), 409. https://doi.org/10.6028/jres.049.044.

(46) Dennis, Jr., J. E.; Moré, J. J. Quasi-Newton Methods, Motivation and Theory. *SIAM Rev.* **1977**, *19* (1), 46–89. https://doi.org/10.1137/1019005.

(47) LaValle, S. M.; Branicky, M. S.; Lindemann, S. R. On the Relationship between Classical Grid Search and Probabilistic Roadmaps. *Int. J. Rob. Res.* **2004**, *23* (7–8), 673–692. https://doi.org/10.1177/0278364904045481.

(48) Govind, N.; Petersen, M.; Fitzgerald, G.; King-Smith, D.; Andzelm, J. A Generalized Synchronous Transit Method for Transition State Location. *Comput. Mater. Sci.* **2003**, *28* (2), 250–258. https://doi.org/10.1016/S0927-0256(03)00111-3.

(49) Verlet, L. Computer "Experiments" on Classical Fluids. I. Thermodynamical Properties of Lennard-Jones Molecules. *Phys. Rev.* **1967**, *159* (1), 98–103. https://doi.org/10.1103/PhysRev.159.98.

(50) Janezic, D.; Orel, B. Implicit Runge-Kutta Method for Molecular Dynamics Integration. *J. Chem. Inf. Comput. Sci.* **1993**, *33* (2), 252–257. https://doi.org/10.1021/ci00012a011.

(51) Andersen, H. C. Molecular Dynamics Simulations at Constant Pressure and/or Temperature. *J. Chem. Phys.* **1980**, *72* (4), 2384–2393. https://doi.org/10.1063/1.439486.

(52) Hoover, W. G.; Holian, B. L. Kinetic Moments Method for the Canonical Ensemble Distribution. *Phys. Lett. A* **1996**, *211* (5), 253–257. https://doi.org/10.1016/0375-9601(95)00973-6.

(53) McDonald, I. R.; Singer, K. Machine Calculation of Thermodynamic Properties of a Simple Fluid at Supercritical Temperatures. *J. Chem. Phys.* **1967**, *47* (11), 4766–4772. https://doi.org/10.1063/1.1701695.

(54) McDonald, I. R.; Singer, K. Examination of the Adequacy of the 12–6 Potential for Liquid Argon by Means of Monte Carlo Calculations. *J. Chem. Phys.* **1969**, *50* (6), 2308–2315. https://doi.org/10.1063/1.1671381.

(55) Torrie, G. M.; Valleau, J. P. Nonphysical Sampling Distributions in Monte Carlo Free-Energy Estimation: Umbrella Sampling. *J. Comput. Phys.* **1977**, *23* (2), 187–199. https://doi.org/10.1016/0021-9991(77)90121-8.

(56) Kästner, J.; Thiel, W. Bridging the Gap between Thermodynamic Integration and Umbrella Sampling Provides a Novel Analysis Method: "Umbrella Integration." *J. Chem. Phys.* **2005**, *123* (14), 144104. https://doi.org/10.1063/1.2052648.

(57) 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.* **1992**, *13* (8), 1011–1021. https://doi.org/10.1002/jcc.540130812.

(58) Laio, A.; Parrinello, M. Escaping Free-Energy Minima. *Proc. Natl. Acad. Sci.* **2002**, *99* (20), 12562–12566. https://doi.org/10.1073/pnas.202427399.

(59) Laio, A.; Gervasio, F. L. Metadynamics: A Method to Simulate Rare Events and Reconstruct the Free Energy in Biophysics, Chemistry and Material Science. *Reports Prog. Phys.* **2008**, *71* (12), 126601. https://doi.org/10.1088/0034-4885/71/12/126601.

(60) *Reviews in Computational Chemistry.*; Parrill, A. L., Lipkowitz, K. B., Eds.; John Wiley: New Jersey, 2015.

(61) Kirkwood, J. G. Statistical Mechanics of Fluid Mixtures. *J. Chem. Phys.* **1935**, *3* (5), 300–313. https://doi.org/10.1063/1.1749657.

(62) Jorgensen, W. L.; Ravimohan, C. Monte Carlo Simulation of Differences in Free Energies of Hydration. *J. Chem. Phys.* **1985**, *83* (6), 3050–3054. https://doi.org/10.1063/1.449208.

(63) Mey, A. S. J. S.; Allen, B. K.; Bruce Macdonald, H. E.; Chodera, J. D.; Hahn, D. F.; Kuhn, M.; Michel, J.; Mobley, D. L.; Naden, L. N.; Prasad, S.; Rizzi, A.; Scheen, J.; Shirts, M. R.; Tresadern, G.; Xu, H. Best Practices for Alchemical Free Energy Calculations [Article v1.0]. *Living J. Comput. Mol. Sci.* **2020**, *2* (1). https://doi.org/10.33011/livecoms.2.1.18378.

(64) Kollman, P. A.; Massova, I.; Reyes, C.; Kuhn, B.; Huo, S.; Chong, L.; Lee, M.; Lee, T.; Duan, Y.; Wang, W.; Donini, O.; Cieplak, P.; Srinivasan, J.; Case, D. A.; Cheatham, T. E. Calculating Structures and Free Energies of Complex Molecules: Combining Molecular Mechanics and Continuum Models. *Acc. Chem. Res.* **2000**, *33* (12), 889–897. https://doi.org/10.1021/ar000033j.

(65) Richmond, T. J. Solvent Accessible Surface Area and Excluded Volume in Proteins. *J. Mol. Biol.* **1984**, *178* (1), 63–89. https://doi.org/10.1016/0022-2836(84)90231-6.

(66) Duan, L.; Liu, X.; Zhang, J. Z. H. Interaction Entropy: A New Paradigm for Highly Efficient and Reliable Computation of Protein–Ligand Binding Free Energy. *J. Am. Chem. Soc.* **2016**, *138* (17), 5722–5728. https://doi.org/10.1021/jacs.6b02682.

# CHAPTER 4. RESULTS AND DISCUSSION

The results presented in this chapter are all derived from computational studies carried out during the course of this thesis. A combination of the techniques described in the previous chapter allowed us to elucidate chemical details at the atomic scale of the catalysis and inhibition processes of the RgpB protease. Mainly, along the whole thesis, QM/MM umbrella sampling[1,2] was used to estimate free energy barriers associated with chemical processes. Furthermore, methods based on classical mechanics were applied to characterize non-reactive processes or phenomena such as binding or conformational processes. A final stage of the thesis was directed towards the exploration of alternative approaches for the estimation of reaction free energies. Specifically, an incursion was made in the implementation of metadynamics in the study of catalytic processes and in the comparison of different mechano-quantum potentials.

## 4.1 ARGININE GINGIPAIN B REACTION MECHANISM

The results presented in this section have already been published in *J. Chem. Inf. Model.* 2021, 61, 9, 4582–4593.

Initially, efforts were concentrated on unraveling the catalytic mechanism by which RgpB carries out protein hydrolysis. For this, we started from the crystal structure deposited by Eichinger, A. and Beisel, H. G. (PDB-ID: 1CVR)[3] of D-Phe-Phe-Arg-chloromethylketone-inhibited gingipain R. The inhibitor initially placed in the active site of the crystallized enzyme was replaced with the Cys-Ala-Tyr-Arg-Thr-Ser-Pro fragment (with acetylated termini) of human pancreatic ribonuclease (UniProtKB-0799+-8)[4] by superimposing the largest number of atoms of the crystallized inhibitor. This fragment has been identified as one of the target sites of RgpB in previous studies[5] and its neutrality helps to avoid the introduction of artifacts into the system. From this base system, we proceeded to study the proteolytic reactivity, starting from the determination of the protonation states of the catalytic dyad (Cys/His) and proceeding with the elucidation of the mechanism itself.

**Protonation States and Structural Stability of the Michaelis Complex of RgpB.** The initial goal was to refine the structure and determine the protonation state of the catalytic dyad His211 and Cys244 in the initial enzyme-substrate complex. For this purpose, 150 ns classical molecular dynamics simulations were carried out over the most possible protonation states of the catalytic diad (Figure 4.1). Thus, three possible protonation states were esxpored: i) R1 that possesses the Cys244 protonated and His211 neutral ii) R2 that possesses the Cys244 deprotonated and the His211 protonated and iii) R3 that possesses both the Cys244 and His211 protonated. The protein was described by means of the AMBERff14SB[6] force field parameters as well as the peptide. A total of 16 to 18 $Na^+$ ions were added,[7] depending on the protonation states of the Cys/His catalytic dyad, in order to neutralize the system. Finally, the system was solvated with a 107.5 $Å^3$ cubic box of water molecules (TIP3P)[8] with a minimum distance of 15 Å between any protein atom and the edge of the box The complete system contains ~116 k atoms. The cutoff limits for short-range nonbonded interactions were 10 Å, and a Particle Mesh Ewald (PME)[9,10] model was used for the long-range interactions. The temperature control was performed using Langevin dynamics[11,12] with a 3 $ps^{-1}$ collision frequency. For all equilibration simulations, the SHAKE algorithm[13,14] was used to constrain light atoms, and the velocity Verlet[15] algorithm was used to update the velocities.

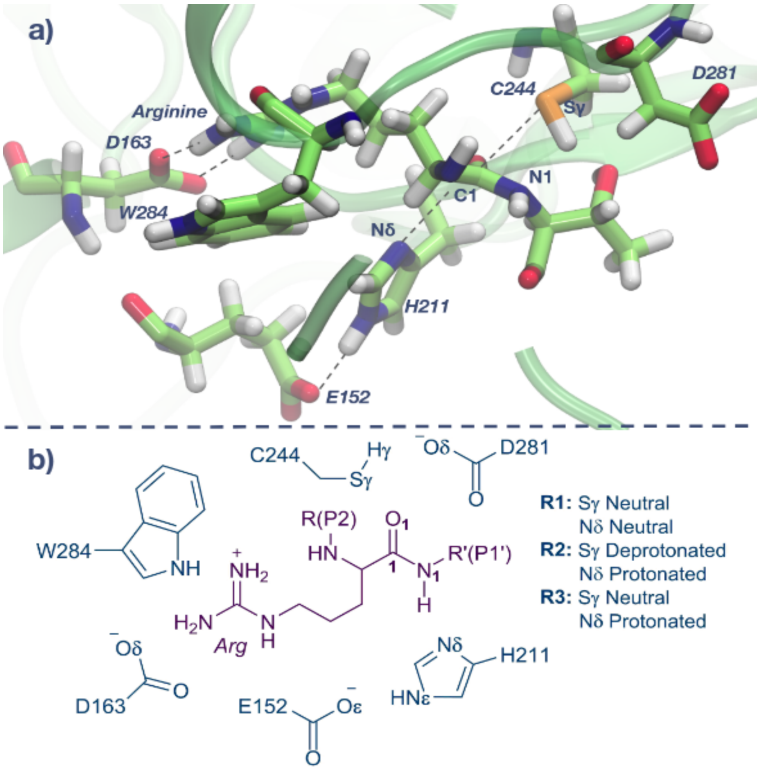


***Figure 4.1.*** *a) Licorice representation of the structure of the active site of RgpB protease of R1 reactant state. b) Sketch of the active site and protonation states variations R1, R2 and R3. Adapted from* [16]

Analysis of the time evolution of the root mean square deviation (RMSD) plots of the protein backbone revealed that all three systems reached equilibrium, and none of the systems exceeded the 2.0 Å threshold of crystal resolution (see Figure 4.2). However, the calculated RMSDs at the active site atoms for the R2 and R3 systems show larger relative standard deviations than for the R1 system (see Figure 4.2). In addition, root mean square fluctuation (RMSF) analysis of the α-carbon atoms of the protein in R2 shows higher fluctuations in the loop between 147 and 159 residues (see Figure 4.2). In light of these results, we conclude that the system with both neutral residues (R1), being more stable, is more viable as a starting point.

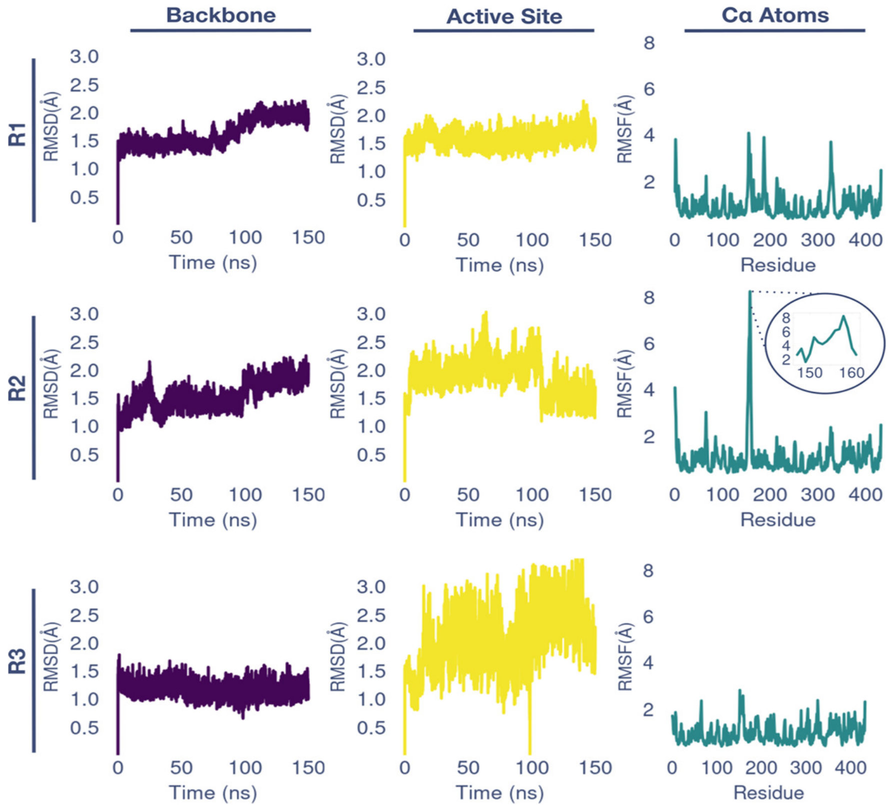


***Figure 4.2.*** *Geometrical analysis of the MD simulations of R1, R2, and R3 systems along the 150 ns of classical MD simulations. Left panels: time evolution of the root-mean-square deviation (RMSD) of protein backbone atoms (Cα, C, N, O). Center panels: time evolution of the RMSD of the active site atoms (heavy atoms of the Glu152, His211, Cys244, Asp281 residues and peptide). Right panels: root-mean-square fluctuation (RMSF) analysis of protein α carbon atoms. The inset in the RMSF plot of the R2 system shows the values of RMSF of the flexible loop 147–159 residues. Taken from* [16]

On further examination, a principal component analysis (PCA) was performed and revealed that the 147-159 loop adopts two possible conformations, open and closed, depending on the protonation of the catalytic dyad. The first major component, PC1, shows such an opening and closing motion, as can be seen in Figure 4.3. The population histogram over PC1 exposes that R3 explores only open conformations, R2 oscillates between open and closed conformations and R1 is restricted to closed conformations. The calculated RMSFs are consistent with a much more fluctuating loop in R2 while R1 and R3 possess a much more rigid loop, albeit in different conformations.

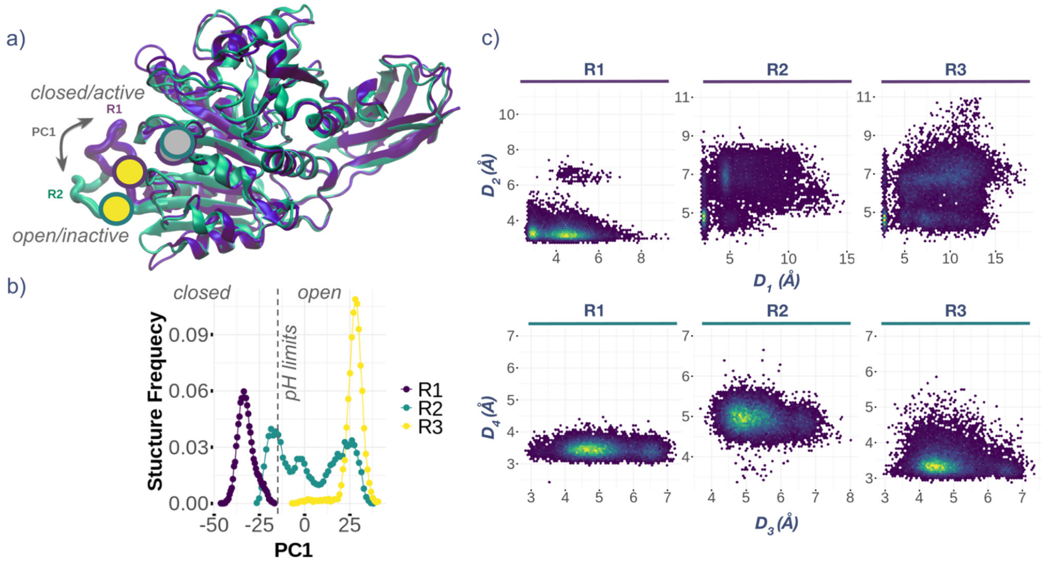


***Figure 4.3.*** *(a) Overlay of the protein structures with the open (R2) and closed (R1) conformations of the 147–159 loop defined by the PC1. Gray and yellow circles represent His211 and Glu152 residues, respectively. (b) Structure frequency of each reactive structure over the PC1 for the open and closed conformations of the 147–159 loop. (c) 2D representations of some distances between the Cys/His catalytic dyad, the peptide, and other residues. D1, [N$\varepsilon$:His211–O$\varepsilon$:Glu152]; D2, [N$\delta$:His211–N1:Pep]; D3, [S$\gamma$:Cys244–O$\delta$:Asp281]; D4, [S$\gamma$:Cys244–O1:Pep]. For D1 and D3, the distances were considered to the closest oxygen atom of the carboxylic group. Distances are in Å. Taken from* [16]

Interestingly, the 147-159 loop contains the Glu152 residue that interacts with and stabilizes the fluctuations of the His211 residue. In open conformations, the catalytic residues Glu152 and His211 are not stably positioned near the active site, so it is valid to suggest that proteolytic activity must be decreased. So far, the only suggested role for His211 is as a proteolysis activating acid.[17] This role would imply a protonated initial state, which according to our simulations results in conformational instabilities. Moreover, simulations with protonated histidine (R2 and R3) showed significantly large fluctuations in some critical distances, these distances are crucial for catalytic activity and therefore rule out the possibility of a protonated His211 in the Michaelis complex.

A population analysis of the most important distances for catalysis revealed that the distance between Oε:Glu152 and Nδ:His211 is influenced by the loop opening. The R2 and R3 systems show unsteady interactions between these two residues (Figure 4.3 panel c)), whereas in R1, Glu152 interacts permanently with His211, as expected in an optimally reactive system. As a consequence, the Nδ:His211-N1:Pep distance remained above 6.5 Å during the R2 and R3 simulations. Whereas, in R1, the distance between these two nitrogen atoms remained stable at an average of 3.2 Å during the MD simulations of the R1 system (see Figure 4.3 panel c)).

Similarly, the interaction between the peptide and the Cys244 residue is also affected by the protonation state of the catalytic dyad. In particular, in system R2, where Cys244 is deprotonated, the average of the position of the sulfur atom appears at a distance of 5.0 Å from the peptide (Sγ:Cys244 and C1:Pep) due to a rotation of the thiolate group, pointing non-conveniently to the opposite direction of the peptide (see Figure 4.3 panel c)). In contrast, the R1 and R3 systems showed an average distance of 3.4 Å between the Sγ:Cys244 and C1:Pep atoms (see Figure 4.3 panel c)). This difference lies mainly in the fact that the Hγ:Cys244 proton in R1 and R3 showed a strong interaction with the O1:Pep atom of the peptide, as revealed by an analysis of the hydrogen bond population. This observed interaction not only places the sulfur close to the electrophilic carbon but also implies that the interaction between residues Asp281 and Cys244, which was proposed to be important for the mechanism,[17] is not frequently observed in any of the simulations. In fact, residue Asp281 was mainly observed interacting with the bulk solvent along our simulations. Previous computational studies[18] support this arrangement by quantifying that the Cys:SH···$^{-}$O-C:Asp interaction is less favorable than Wat:OH···$^{-}$OC:Asp.

In all, our simulations pointed to the His211 and the Cys244 residues in their neutral states (R1) as the most stable, ergo plausible, precatalytic state. Other protomers (R2 and R3) proved to be unstable and thus unlikely as catalytic structures. The cysteine protease legumin, which shows a similar spatial distribution of the catalytic Cys/His dyad, was proposed to works from the same protonation states, according by Elsässer and co-workers.[19]

Finally, the calculated pKa values for Cys244 and His211 residues in the open conformation are 6.31 and 9.92, respectively. These values closely match the pKa experimentally estimated for RgpB[20] and coincide with the pH limits from which catalytic activity decreases dramatically because of structural instability. In light of our findings, such structural instability may be related in a certain way with the open/closed conformations of the loop.

**Proteolysis Reaction Mechanism.** Once it was concluded that R1 is the most probable pre-catalytic state, we proceeded to elucidate the mechanism of the proteolysis reaction. For this purpose, the free energy surfaces of the different reaction pathways were calculated by umbrella sampling[2] at the PM6/MM level (QM region represented in Figure 4.4), with the same parameters employed in the classical simulations for the MM region. Each window of the umbrella sampling had a relaxation time of 5 ps and a sampling time of 25 ps. The force constant used for each window was 580 $\text{kcal} \cdot \text{mol}^{-1} \cdot \text{Å}^{-2}$, and the window width was between 0.05 and 0.1 Å depending on the reaction coordinate. The number of windows were selected to ensure a correct overlapping of sampling along the colective variable. The collective variables used in the umbrella sampling calculations are described below. From the free energy surfaces obtained at the PM6/MM level, the potential energy of the critical points was corrected by single point calculations at the PBE+D3(BJ)/MM level using a 6-311+G** basis set. The geometries used for the corrections were obtained by optimizing a representative transition state structure and subsequently reaching minima by intrinsic reaction coordinate (IRC) calculations. All umbrella sampling simulations were carried out in NVT conditions at 310 K.

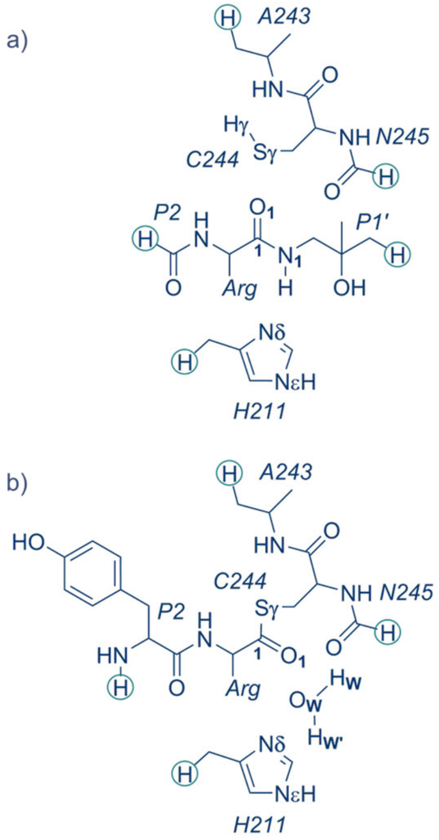


**Figure 4.4.** *Schematic representation of the region treated quantum mechanically to explore the mechanism of (a) acylation stage and (b) deacylation stage (link atoms added to treat QM–MM frontiers depicted in circles). Taken from* [16]

Starting from protonated Cys244 and deprotonated His211 (R1), several mechanisms were evaluated, see (Figure 4.5). The free energy profiles of the most feasible reaction mechanism are depicted in Figure 4.6 while for the rest of the mechanisms are presented in Figure 4.7. The stationary point structures for the energetically most favorable mechanism are depicted in Figure 4.8.

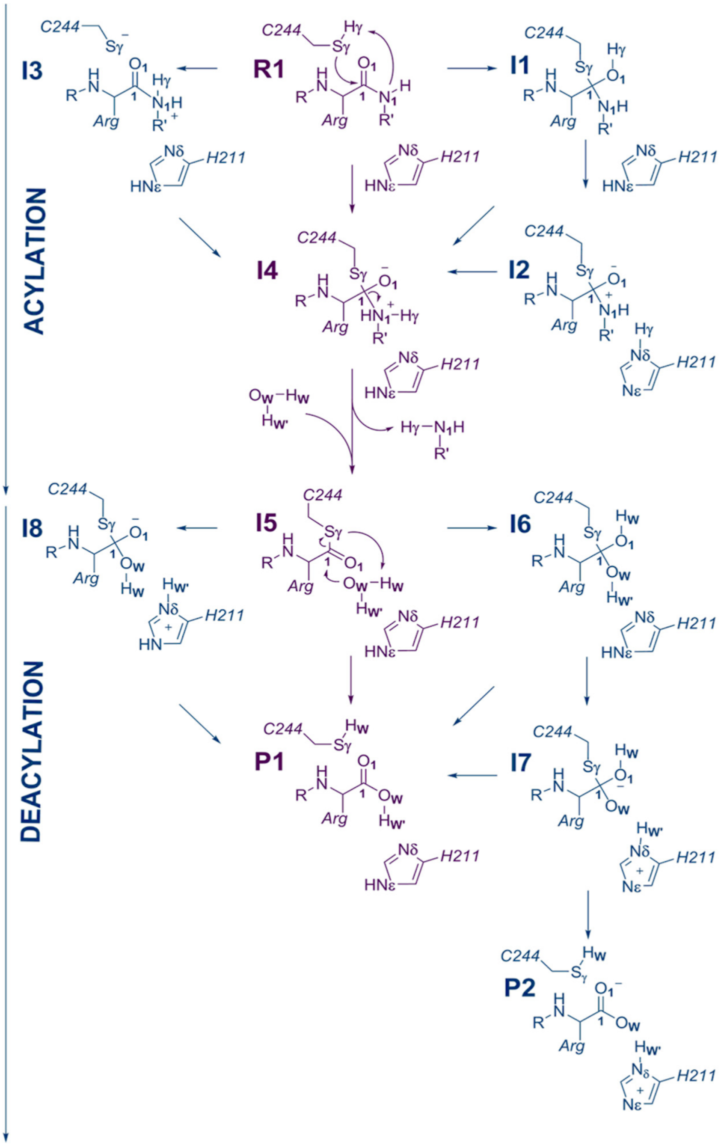


***Figure 4.5.*** *Studied plausible reaction paths for the proteolysis catalyzed by the RgpB gingipain. Most favorable mechanism according with our QM/MM calculations is depicted un purple. Taken from* [16]

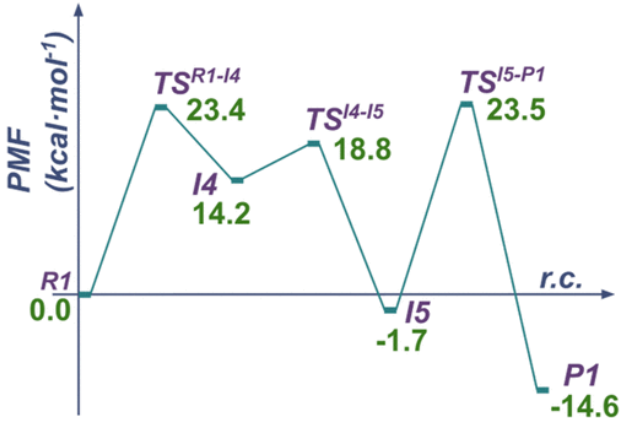


***Figure 4.6.*** *PBE+D3(BJ):PM6/MM free energy profile of the proteolysis catalyzed by the RgpB gingipain. Taken from* [16]

The first possibility considered consisted of proton transfer from the Sγ:Cys244 atom to the O1 atom of the peptide (O1:Pep) together with nucleophilic attack of the Sγ:Cys244 atom to the C1:Pep atom, obtaining intermediate I1. This process transits through an activation barrier of 24.5 kcal · mol$^{-1}$ associated with the TS$^{R1\text{-}I1}$ transition state (see figure 4.7). The intermediate I1 formed is located at 15.6 kcal · mol$^{-1}$ above R1 on the free energy surface. From I1, the intermediate I4 can be reached in two ways, i) by transferring the proton from the O1:Pep atom to the Nδ:His211 atom (I1 → I2), which would serve as a halfway intermediate to transfer the proton to the N1:Pep atom in a subsequent step (I2 → I4) or ii) a direct transfer from the O1:Pep atom to the N1:Pep atom (I1 → I4). The first option shows free energies of 29.3 kcal · mol$^{-1}$, 28.7 kcal · mol$^{-1}$ and 31.0 kcal · mol$^{-1}$ associated with TS$^{I1\text{-}I2}$, I2 and TS$^{I2\text{-}I4}$, respectively, making this mechanism unfeasible due to the high energies linked to their transition states. Similarly, the second option, as expected, carries a high energy penalty, 46.5 kcal · mol$^{-1}$ over R1 as it proceeds through a four-membered transition state (TS$^{I1\text{-}I4}$). Although the formation of I1 is energetically possible, proceeding from this point to the acylenzyme is infeasible and it is unlikely that the mechanism proceeds by any of the options described.

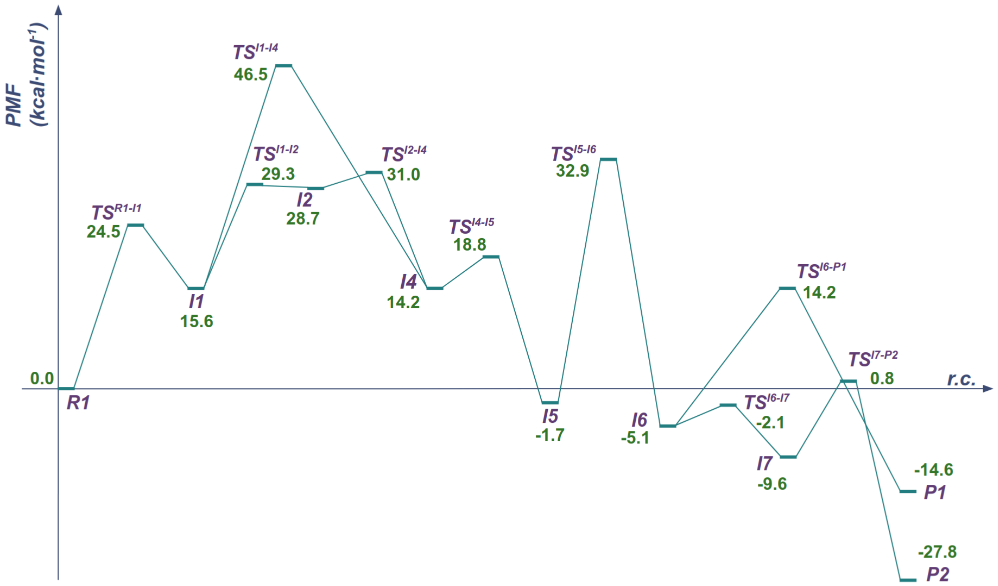


**Figure 4.7.** *PBE+D3(BJ):PM6/MM free energy profile of the alternative reaction pathways computed for the proteolysis catalyzed by the RgpB gingipain. Taken from* [16]

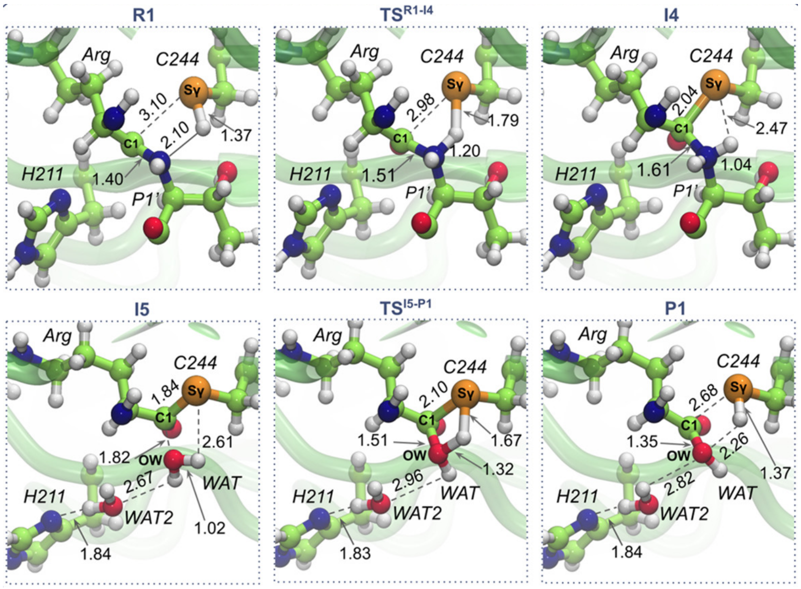


**Figure 4.8.** *PM6/MM optimized structures of the key states along the most favorable proteolysis reaction mechanism: R1, $TS^{R1\text{-}I4}$, I4, I5, $TS^{I5\text{-}P1}$, and P1. Distances are presented in Ångstroms. Adapted from* [16]

Another mechanism considered, starting from R1, involves proton transfer from the Sγ:Cys244 atom to the N1:Pep atom (R1 → I3) and subsequent nucleophilic attack of the Sγ:Cys244 atom to the C1:Pep atom (I3 → I4). However, the putative intermediate I3 could not be located or observed on any of the calculated free energy surfaces. On the other hand, the analogous concerted reaction mechanism from R1 to I4 proceeds through a favorable activation energy of 23.4 kcal · mol$^{-1}$. This attack of the Sγ:Cys244 atom on the C1:Pep atom is promoted by the oxyanionic hole of RgpB, which comprises the nitrogen protons from the backbone of residues Gly212 and Cys244. The intermediate obtained, I4, has an associated free energy of 14.2 kcal · mol$^{-1}$ with respect to R1. This energy is explained by structural instability of I4, reflected in the longer than usual bond distances N1:Pep-C1:Pep and Sγ:C211-C1:Pep (1.61 and 2.04 Å, respectively). In fact, the decomposition of I4 into intermediate I5 by cleavage of the N1:Pep-C1:Pep-C1:Pep peptide bond follows a low relative barrier of 4.6 kcal · mol$^{-1}$. The acylenzyme, I5, is localized to -1.7 kcal · mol$^{-1}$ of R1 on the free energy surface. This I5 intermediate is structurally and energetically similar to the analogous structure reported by Elsässer et al.[19]

From the mechanisms explored for the acylation step (R1 to I5), the most favorable reaction pathway takes place in two steps, via a zwitterionic stable intermediate, I4 (see Figure 4.6). The energy obtained for the rate-determining step is 23.4 kcal · mol$^{-1}$, associated with the TS$^{R1\text{-}I4}$ transition state. This activation barrier is congruent with the reported experimental data, 22.8 kcal · mol$^{-1}$.[21] The role of the N1:Pep atom as an activating base has been previously reported in the catalytic mechanism of legumin protease.[19] Such a role of the substrate explains the inability of compound E-64[22] to inhibit this family of enzymes, lacking a properly disposed group that can play the role of a base. In fact, the presence of a properly disposed base group must be carefully considered when designing covalent inhibitors for arginine gingipain B. Also, due to the values of the activation and reaction free energies of the R1 → I1 step, similar to those of the R1 → I4 step, considering new inhibitors following this alternative mechanism would be fruitful.

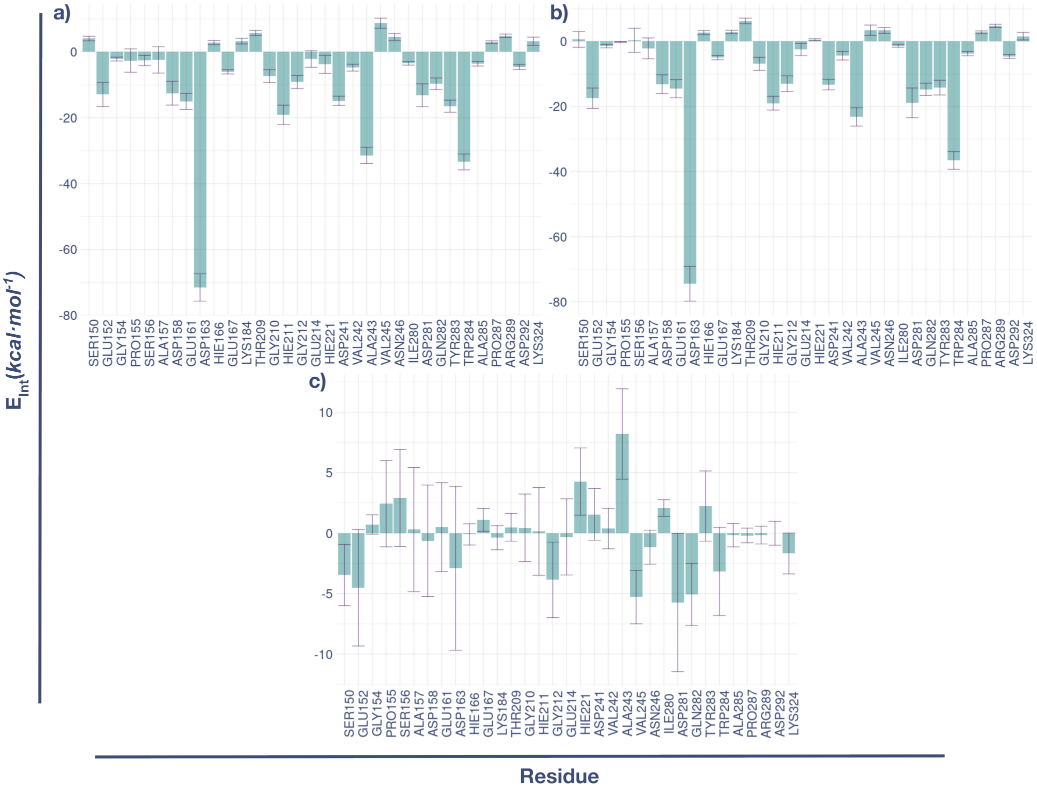


***Figure 4.9.*** *Main average interaction energies (electrostatic plus Lennard-Jones) between peptide and the protein residues along 1 ns of PM3/MM MD simulations at a) the R1 reactant state, b) the* $TS^{R1\text{-}I4}$ *transition state and c) the difference between* $TS^{R1\text{-}I4}$ *and R1. Adapted from* [16]

In addition to the free energy surfaces, the evolution of the interaction energies between the residues of the protein and peptide along the first step was also calculated (see Figure 4.9). According to the results, residues Asp163 and Trp284 interact strongly with the guanidinium group of the arginine of the peptide through salt bridge and π-stacking interactions, respectively, which derives in the high selectivity of RgpB for arginine residues. In addition, residues His211 and Glu152 also showed highly favorable peptide interactions. Revealing a crucial structural/electrostatic role of these residues in peptide-enzyme interaction and reactive complex formation. Indeed, the oxyanionic hole residues, Cys244 and Gly212, exhibited lower interaction energies than residues His211 and Glu152.

Classical molecular dynamics simulations revealed that two water molecules occupy the vicinity near to the new Sγ:Cys244-C1:Pep bond in the space generated by the release of the first peptide fragment. These molecules occupy that space only if the first fragment is released and are the ones that will then participate in the hydrolysis of adduct I5.

As for acylation, several mechanisms were considered. Initially, on the basis of a previous study,[19] we considered the attack of a water molecule on the C1:Pep atom in concert with proton transfer to the O1:Pep atom for the formation of a geminal diol (I5 → I6). While the intermediate I6 is stable (-5.1 $\mathrm{kcal \cdot mol^{-1}}$), as in the previous study, in the case of RgpB it has a much higher activation energy, 32.9 $\mathrm{kcal \cdot mol^{-1}}$. This activation energy rules out the possibility of the reaction proceeding via I6, even if after I6 the products are easily reachable via a low-barriers mechanism (I6 → I7 → P2, Figure 4.7).

Alternatively, it was explored the possibility that residue His211 acted as a base, deprotonating the water molecule, assisting the attack on the C1:Pep atom to reach intermediate I8. However, I8 was shown not to be a minimum over the free energy surfaces evaluated, which finally ruled out the possibility that His211 plays a covalent role during catalysis. In fact, a inspection of the I5 classical molecular dynamics revealed that the His211 role is disposing, via a water-mediated hydrogen bond, one of the nucleophilic water protons oriented directly toward the Sγ:Cys244 atom.

Based on this observed arrangement, direct proton transfer from the water molecule to the Sγ:Cys244 atom was evaluated along with nucleophilic attack of the water molecule on the C1:Pep atom. The free energy surface indicated these two processes occur in concert across a 23.5 $\mathrm{kcal \cdot mol^{-1}}$ barrier associated with the $\mathrm{TS^{I5\text{-}P1}}$

transition state, making this the most favorable reaction mechanism for the hydrolysis of the I5 adduct. The energy associated with the deacylation process could dictate the rate limiting step of the overall reaction. However, considering the similarity between the energies of $TS^{I5\text{-}P1}$ and of the first transition state of the acylation process, $TS^{R1\text{-}I4}$, both steps should contribute strongly to the kinetics of the overall reaction. The relative free energy of the reaction product, P1, is -14.6 $kcal \cdot mol^{-1}$ compared with R1, resulting in an exergonic process. Although no experimental reference is available for the deacylation step, some of the studies previously performed in our group show similar free energies in analogous processes.[23,24]

The geometries of the transition states of the kinetically relevant steps (the rate determining steps of the acylation and deacylation stages) were optimized at PBE+D3(BJ)/MM level (Figure 4.10). A comparison of these structures with those obtained at the lowest level of theory (PM6/MM, Figure 4.8) shows structural similarities that support the robustness of our conclusions based on PM6/MM method.

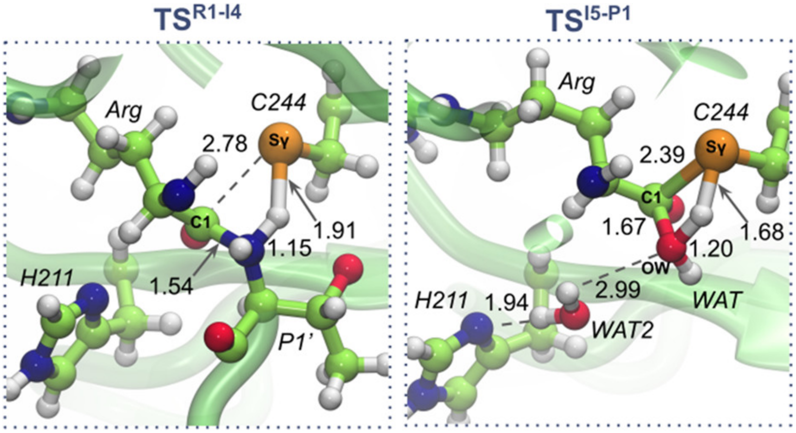


**Figure 4.10.** *PBE+D3(BJ)/MM optimized structures of the rate limiting transition states of the acylation ($TS^{R1\text{-}I4,}$ left panel) and deacylation ($TS^{I5\text{-}P1}$, right panel) stages in the reaction of proteolysis catalyzed by the RgpB gingipain. Adapted from* [16]

## 4.2 ARGININE GINGIPAIN B COVALENT INHIBITION

The results presented in section 4.2 have already been formatted as manuscript and submitted for publication.

The second stage of the thesis was focused on studying the inhibition of RgpB gingipain as an attempt to search for potential treatment for Alzheimer's disease. For this purpose, from the patent published by Dominy and coworkers (Patent Aplication PCT/US2016/061197), a pool of molecules representing a good spectrum of the chemical space of the reported compounds was selected (Figure 4.11). As described in the introduction, the compounds developed and published by Dominy and coworkers,[25] showed strong neuroprotective activity in mice. This fact postulated this family of compounds as potential treatments against Alzheimer's disease and positions them in the focus of drug development research. These compounds were classified as irreversible inhibitors with a potent mean inhibitory concentration ($IC_{50}$) of $<50$ pM.[25] As irreversible inhibitors, and in order to enrich progress in the development of treatment portfolios, the correct characterization of their activity on both the description of the non-covalent binding process and the covalent process is essential. For this purpose, we employed MM/PBSA and alchemical transformations for the quantification of the binding free energies, while for the evaluation of the free energy surfaces of the covalent binding mechanism we used the same methods and simulation parameters as in the study of the proteolysis reaction catalytic mechanism (section 4.1).

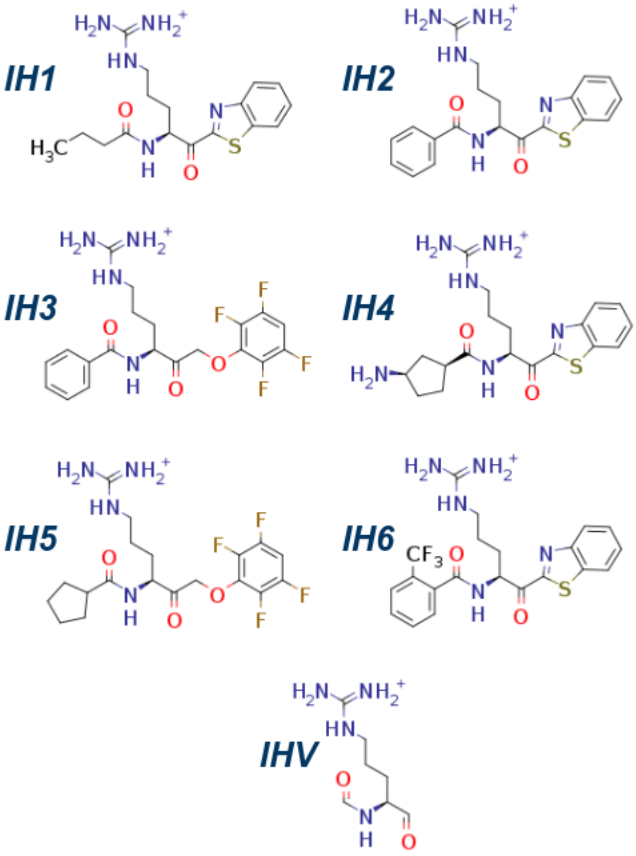


*Figure 4.11.* *Sketch of the studied irreversible inhibitors of RgpB (IH1-IH6) and of the virtual inhibitor (IHV) used as common point for the alchemical transformations.*

**Binding affinities ($\Delta G_{bind}$) and ligand-receptor interaction profiles.** In order to study the non-covalent interaction between inhibitors and RgpB enzyme, 250 ns of classical MD simulations were initially run in the inhibitor:receptor non-covalent reactant complex per inhibitor. None of the systems, in the presence of the corresponding inhibitor, exhibited significant changes in the protein structure (all backbone RMSDs < 2 Å on average, Figure 4.12).

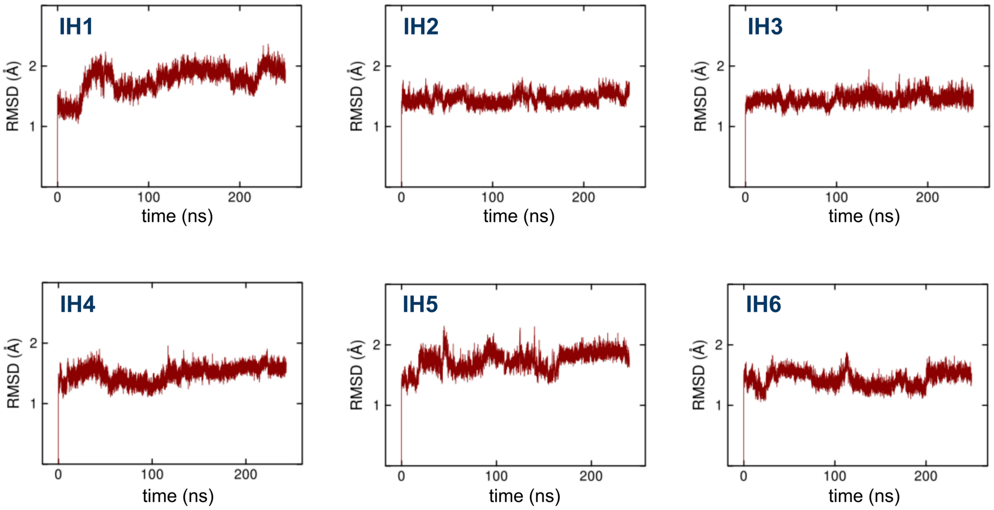


***Figure 4.12.*** *Time evolution of the root-mean-square deviation (RMSD) of protein backbone atoms (Cα, C, N, O) of the RgpB gingipain in complex with the six selected inhibitors (IH1 to IH6).*

Starting from the equilibrated structures of the classical MD simulations, we proceeded to estimate the binding free energies for each of the compounds. For this purpose, we computed the binding affinity energies by MM/PBSA calculations supplemented with a correction term to the entropic contributions based on the interaction entropies proposed by Duan et al.[26] (Table 4.1). Qualitatively speaking, compounds IH3 and IH4 proved to produce the most stable inhibitor-enzyme complexes. It should also be noted that IH3 is the one that shows the most favorable binding energy in both entropic and enthalpic terms. On the other hand, inhibitor IH2 presented a lower binding enthalpy and a higher -TΔS term. We emphasize that the values computed by means of MM/PBSA and interaction entropies have no quantitative meaning and are purely used to analyze qualitatively the enthalpic and entropic contributions of the binding processes.[27] Thus, keeping in mind the inherent uncertainty associated to the MM/PBSA method, alchemical transformations[28] were employed to compute the differences between the binding free energies of every inhibitor and the virtual inhibitor IHV ($\Delta\Delta G_{bind\text{-}TI}$) in order to obtain more precise and quantitatively meaningful inhibitor-enzyme affinities.

***Table 4.1.*** *Binging free energies for the studied inhibitors obtained over the classical MD simulations using MM/PBSA and alchemical transformations (thermodynamic integration) methods. Values derived from the later ($\Delta G_{bind\text{-}TI}$) are reported as relative to the virtual inhibitor IHV. All values are in kcal · mol$^{-1}$.*

| **Inhibitor** | $\Delta H_{bind\text{-}MM/PBSA}$ | $-T\Delta S_{bind\text{-}MM/PBSA}$ | $\Delta G_{bind\text{-}MM/PBSA}$ | $\Delta\Delta G_{bind\text{-}TI}$ |
|---|---|---|---|---|
| **IH1** | -57.3 | 37.0 | -20.3 | 0.20 |
| **IH2** | -54.2 | 50.6 | -3.6 | 5.42 |
| **IH3** | -69.4 | 27.3 | -42.1 | 0.55 |
| **IH4** | -67.2 | 28.2 | -39.0 | 1.99 |
| **IH5** | -63.3 | 27.8 | -35.5 | 2.55 |
| **IH6** | -60.9 | 33.1 | -27.8 | 3.31 |

In order to obtain accurate values and to analyze from a more reliable ranking, the differences between the binding free energies of every inhibitor and the virtual inhibitor IHV ($\Delta\Delta G_{bind\text{-}TI}$) were computed by alchemical transformations.[28] For this purpose, the 3-step Amber Thermodynamic Integration protocol ("decharge-LJ-recharge" protocol) was used.[29] Each transformation was carried out in ten windows equally distributed throughout the λ range (0-1). In each window 5 ns of sampling was performed. Only the atoms appearing/disappearing during the transformation were included in the soft-core region. All the simulations were performed at NVT ensemble at 310 K, starting from the volume equilibrated structures of the classical MD trajectories.

The results of the alchemical transformations show a similar trend to the one obtained from MM/PBSA calculations, with the only exception of IH1, previously ranked as the second weakest, is repositioned in first place with practically the same $\Delta\Delta G_{bind\text{-}TI}$ as IH3. Same as the MM/PBSA calculations, IH4 ranks after IH3 as one of the most potent candidates. Likewise, IH2 is the inhibitor with the lowest affinity to the enzyme with a difference of 5.2 kcal·mol$^{-1}$ with respect to IH1. Henceforth, IH3 can be considered as the reference inhibitor given the agreement between the binding free energies estimated by both methods.

A contact frequency map allowed us to analyze the differences in the interaction patterns/profiles of each inhibitor, which can be complemented with the analysis of the averaged enzyme-inhibitor interaction energies decomposed by residue (Figure 4.13). Figure 1a shows the relative contact frequencies with respect to the inhibitor with the highest affinity, IH3. Thus, positive values represent a higher contact frequency than in IH3 while negative values represent a lower contact frequency than in IH3. In general, most of the inhibitors present a similar interaction profile with the receptor, which is in agreement with the differences observed between the binding free energies ($\Delta\Delta G_{\text{bind-TI}}$) computed by alchemical transformations. However, it can be observed from the contact map that three inhibitors (IH2, IH4 and IH6) show a significantly lower contact frequency with residues Ser213 and Glu214 (see Figure 4.13).

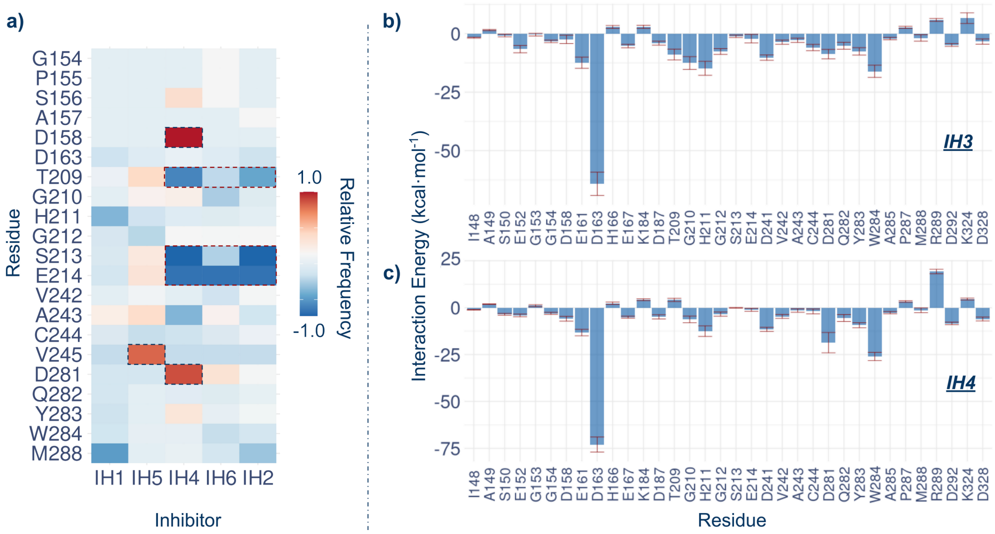


***Figure 4.13.*** *a) Relative frequency contacts map between RgpB residues and the inhibitors. IH3 was used as frequency reference. A contact was counted if the distance*

*between atoms was < 4.0 Å. The cells corresponding to the interactions that differ the most from the IH3 inhibitor are highlighted with dashed-line borders. Panels b) and c) show averaged interaction energies (electrostatic plus Lennard–Jones) between residues of RgpB and IH3 and IH4, respectively.*

A visual inspection revealed that for the case of inhibitors IH3 and IH5 this interaction corresponds to a hydrogen bond between the HN:Glu214 and the F1 atom of the inhibitor. In the case of the interaction with Ser213 residue does not reach a hydrogen bond due to the angle of the atoms involved (angle around 59±16 degrees) (Figure 4.14). Thus, for those inhibitors that instead of the tetrafluorophenyl substituent have the benzothiazole ring, the interactions with Ser213 and Glu214 are not present. However, IH1 manages to interact with HN:Glu214 via the nitrogen atom of the bicyclo ring. Although no preferences in selectivity of the groups around arginine moiety have been reported in RgpB that affect the catalytic capacity, a higher affinity for hydrophobic substituents has been shown, mainly aromatic. However, these results suggest that hydrogen bonding groups are necessary to stabilize these aromatic substituents within the active site and can be optimized to achieve better affinity to the receptor.

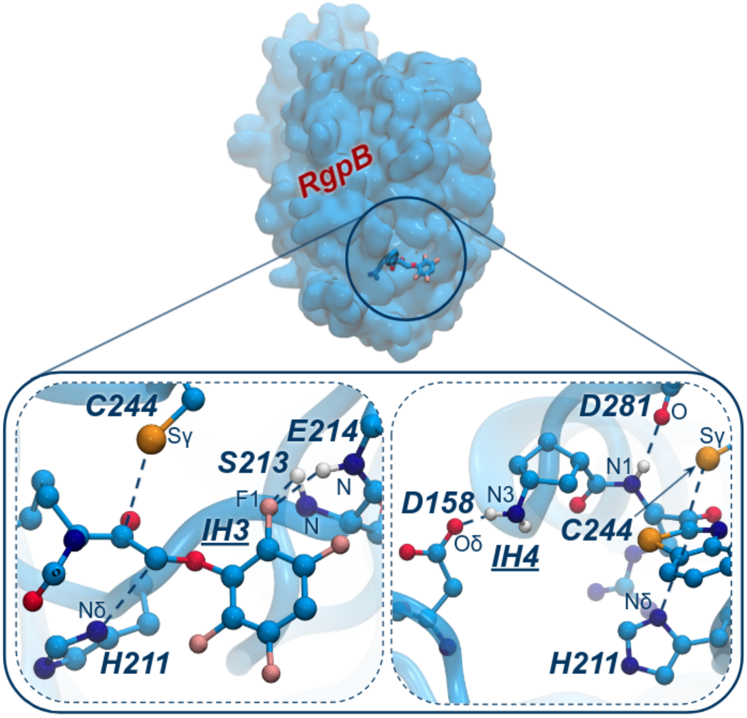


***Figure 4.14.*** *Top panel: Representation of RgpB gingipain (surface) with non-covalently bounded IH3 inhibitor, represented as sticks. Bottom panels: CPK representation of the IH3 (left) or IH4 (right) inhibitors in the binding pocket of RgpB. Key interactions are highlighted with blue dashed lines.*

It is also worth highlighting the interaction observed between compound IH4 and residues Asp158 and Asp281 (see Figure 4.14). Namely, IH4 is the only one of the compounds studied that presents a hydrogen bond donor at the nitrogen substituent group of the arginine. As a consequence, HN3:IH4 interacts with Oδ:Asp158 (Figure 4.14). As a result, this inhibitor is repositioned in such a way that it forms an extra hydrogen bond interaction between the HN1:IH4 and O:Asp281 atoms. Although this interaction would be possible in all inhibitors, it was only observed in IH4, suggesting that it responds to the conformation adopted due to the interaction with residue Asp158. Other differences in the contact map, such as those observed in residues Thr209 or Val245, are less specific and are the consequence of a particular physical proximity during the simulations.

Finally, in order to energetically characterize the favorable observed contacts between the selected inhibitors (IH3 and IH4) and the RgpB residues, the interaction energies decomposed by residue were calculated (Figure 4.13 b) and c)). We can observe the determinant role played by residues Asp163, Trp284 and His211 in arginine binding, a conclusion that was predicted from structural analysis and our previous report on the proteolysis reaction catalyzed by RgpB. This result is not surprising given the high selectivity of RgpB to arginine residues. On the other hand, it can be observed that the interaction with residues Ser213 and Glu214 is always favorable. In the same way, the interaction between IH4 and residues Asp158 and Asp281 shows a lower energy, being remarkable the one with Asp281. In contrast, residues Lys184, Arg289 and Lys324 show an unfavorable interaction energy with all compounds. Although this may be a starting point for future optimizations, it derives from the spatial proximity (without contact) between the charges of these residues and the positive charge of the guanidinium group of the arginine moiety of the inhibitors.

To sum up, our results suggest that the optimization should be focused on compounds presenting hydrogen bond acceptor groups capable to interact with HN:E214 and HN:Ser213. Moreover, hydrogen bond donor groups interacting with Asp158 and Asp281 seems to enhance the affinity with the receptor. On the other hand, the guanidinium group present in all inhibitors and in the natural substrate should be preserved as has been shown to be essential for recognition by RgpB enzyme.

**Covalent binding chemical step.** In order to study the reaction mechanism by which these inhibitors covalently bind to the enzyme, QM/MM MD simulations were performed to generate the full free energy landscapes of the most plausible mechanisms. Given the similarity between the inhibitors, the reactivity study was carried out using only the inhibitor IH3, which showed the most favorable binding energy.

Initially, and analogously to the proteolysis reaction catalyzed by the enzyme in the presence of wild-type substrate, Cys244 is protonated and requires activation to attack the carbonyl group in the reactant state (R in Figures 4.15 a) and b), and Figure 4.16 b)). We also considered the possibility that the mechanism proceeded starting from the deprotonated Sγ:Cys244 form (mechanism $R_{S(-)} \rightarrow P_{S(-)}$ in Figure 4.15 c)). However, several attempts (with and without restrains) to obtain stable reactive structures were unsuccessful. In all simulated cases, the negative charge on Cys244 resulted in deformations of the system to chemically unviable structures.

The family of inhibitors reported does not possess a clear reactive moiety, such as a Michael acceptor or an epoxide group. Instead, they only possess the carbonyl oxygen atom O1:IH3 as a possible acceptor/activator for the cysteine residue. This possibility was previously hypothesized[16] in light of the small difference between the free energy barriers shown by the wild-type being activated by either the peptide nitrogen (more favorable mechanism by ~1.1 kcal·mol$^{-1}$) or carbonyl oxygen of the substrate. Other possible bases were evaluated by calculating the population of hydrogen bonds between Hγ:Cys244 and other possible proton acceptors. However, the interaction with O1:IH3 was shown to populate more than 50% of the time, while other possible bases, such as the Asp281, interacted less than 1% with it, making them poor base candidates. With the carbonyl group as the only plausible activator of Cys244, the mechanisms are limited to whether the attack of Sγ:Cys244 on C1:IH3 and the transfer of Hγ:Cys244 from Sγ:Cys244 to O1:IH3 occur directly or mediated by a water molecule (Figure 4.15 a) and b)).

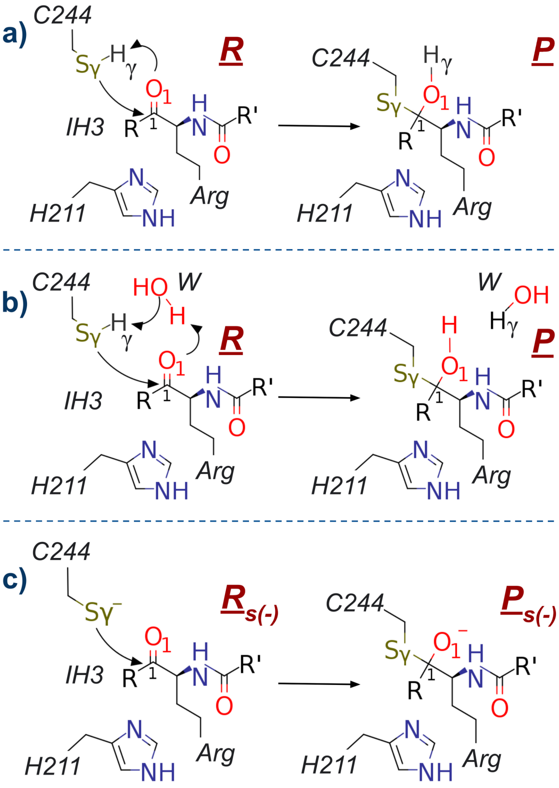


***Figure 4.15.*** *Schematic representation of the considered mechanisms of covalent binding between IH3 and RgpB. Sketch of the reaction mechanism in which the Hγ:Cys244 transfer to the O1 atom of the inhibitor occurs a) directly or b) mediated by a water molecule. c) Representation of the reaction mechanism starting from the deprotonated Sγ:Cys244 form. For IH3, R and R' correspond to phenyl and 2,3,5,6-tetrafluorophenoxy substituents, respectively.*

As revealed by the computed profiles, both mechanisms, the one in which the Hγ:Cys244 is directly transferred to the O1:IH3 atom (mechanism a in Fig. 4.15), or the one mediated by a water molecule (mechanism b in Fig. 4.15), proceed through a single concerted step. In the case of the direct mechanism, that was explored using the distance Sγ:Cys244-C1:IH3 and the difference between distances [Sγ:Cys244-Hγ:Cys244] - [O1:IH3-Hγ:Cys244] as collective variables to describe the process, the activation free energy associated to the transition state (TS, see Figure

4.16 a)) is 22.8 kcal·mol$^{-1}$ with respect to the reactants. A value that is very close to that previously computed for the natural substrate, 23.5 kcal·mol$^{-1}$.[16] However, in the case of the transition state mediated by a water molecule (TS$_W$), that was explored using the distance Sγ:Cys244-C1:IH3 and the difference between distances [Sγ:Cys244-Hγ:Cys244] - [O:W-Hγ:Cys244] as collective variables to describe the process, the corresponding activation energy is 14.2 kcal·mol$^{-1}$. This result suggests this mechanism as the most viable for the covalent binding reaction with a significantly lower free energy barrier compared to that of the wild-type substrate. The obtained product (P) is located at -12.1 kcal·mol$^{-1}$ with respect to the reactants, thus giving an irreversible exergonic reaction. In order to verify the mechanism obtained at PM6/MM level, R, TS$_W$ and P were fully optimized at a higher level of theory, PBE+D3(BJ)/MM. The high-level optimized structures, presented in Figure 4.16, match the reaction pathway predicted at lower level, verifying the mechanism deduced from the PM6/MM FES.

As mentioned above, the mechanism starting from the deprotonate Cys244 (mechanism c in Figure 4.15) was not explored because the active site was deformed along the classical MD simulations and no properly reactive structures were obtained.

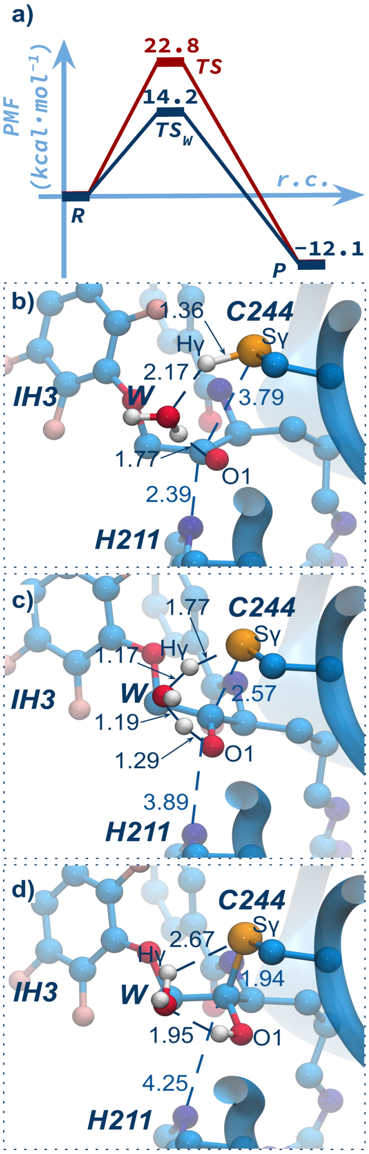


**Figure 4.16.** *a) Free energy profiles computed for the inhibitory covalent binding mechanisms of RgpB by IH3 at the PBE+D3(BJ):PM6/MM. The direct proton transfer mechanism from Cys244 to O1 atom of IH3 is depicted in red while the proton transfer mechanism mediated by a water molecule is depicted in blue. PBE+D3(BJ)/MM optimized structures of the critical points b) R, c) $TS_W$ and d) P along the most likely*

*covalent binding mechanism of IH3 inhibiting RgpB catalytic activity. Selected distances are in Å.*

In light of these results, it is remarkable to highlight the ability of RgpB to react without the presence of highly reactive groups. Common reactive groups such as Michael acceptors or epoxides deal with selectivity issues due to their reactivity with other undesired targets. The unusual carbonyl warhead enables to exploit the design of covalent inhibitors, with high potency, without having to face problems of selectivity. These features qualify RgpB as a pharmacological target that promises effective treatments by the use of these kind of inhibitors without side effects for the treatment of Alzheimer's disease.

## 4.3 FULLY DFT-BASED FREE ENERGY SURFACES BY MTD SIMULATIONS

The results presented in section 4.3 have already been formatted as manuscript and submitted for publication.

Calculating a free energy surface at the full DFT level is often a resource-intensive task, especially in biochemical systems where quantum regions often contain hundreds of atoms and sometimes include metals. For this reason, exhaustive sampling methodologies such as umbrella sampling require variations to reduce the amplitude of the phase space to be sampled when performed with DFT methods. The most common variations are usually path-based methods,[30–35] in which the multidimensionality of the surfaces is projected onto a single dimension that exclusively describes the process of interest, thus reducing the number of windows needed to estimate the process barriers.

Hitherto, the reactivity descriptions of the reactivity and inhibition of RgpB have been carried out in the present thesis using umbrella sampling methods based on hybrid QM/MM potentials, using the PM6 semi-empirical method to describe the QM region and subsequently corrected using a DFT method, in this case PBE-D3(BJ). Thus, the scheme calculates the entropic and zero-point contributions at the semi-empirical level and the potential energy at the DFT level.

This scheme is quite robust and has been widely used (or variants) in reactivity studies.[23,24,36–38] It optimally balances accuracy and computational cost. However, it should be kept in mind that on some particular occasions non-potential terms calculated at the semi-empirical level may lead to misestimations. It should be noted that as in all cases, the choice of method depends largely on the nature of the system to be modeled, and a method with a higher level of theory does not necessarily always lead to more reliable results.

Alternatively, there are also additional methodologies that allow to reduce the exhaustiveness of the sampling by limiting to evolve along the lowest free energy reaction path in a single simulation. This is the case of the metadynamics simulations that were described in detail in the third chapter. Metadynamics reduces the number of net steps required to obtain a reasonable free energy profile.[39] At this stage of the thesis the focus became methodological in favor of learning the theory and implementation of fully DFT metadynamics simulations.

The system selected for the metadynamic studies was the ATPase/helicase Prp2. Because of the nature of the system (highly charged species in the active site including magnesium cations and phosphorous atoms), this is an example where severe limitations can be found if the reactive part of the system (the QM region) is treated with a semiempirical method during the conformational sampling. From the biological point of view, this DEAH-box family member enzyme modulates, along the RNA splicing cycle, the critical rearrangement of the spliceosome complex to its catalytically competent form.[40–43] Briefly, Prp2 forms a complex with the pre-mRNA and hydrolyzes ATP to perform a stepwise translocation of the pre-mRNA to its 3' end. The core architecture of Prp2 is conserved among other spliceosomal DExH-box ATPases/helicases (i.e. Prp16, Prp22 and Prp43).[44] In all of these enzymes, a single-stranded RNA, upon binding in a tunnel formed at the interface of the RecA1, RecA2 and HB, OB, WH domains, allosterically activates ATPase function. Recent Cryo-EM and X-ray resolved structures of Prp2 in the presence of RNA and ADP/ATP,[44] provided valuable information to formulate mechanistic hypotheses about Prp2 helicase activity. However, the molecular mechanism of ATP hydrolysis of Prp2 with atomic-level details and its RNA-driven allosteric regulation remains elusive. Although a complete mechanistic understanding of pre-mRNA splicing is linked to the molecular details of ATPase/helicase function underlying critical steps in spliceosome remodeling.

**Enzymatic mechanism of ATP hydrolysis in Prp2.** In order to elucidate the mechanism of ATP hydrolysis catalyzed by Prp2, we employed metadynamic simulations at hybrid quantum/classic (QM/MM) theory level, based on DFT-BLYP-D3 for the QM part and on AMBER force fields for the MM part.[45–47] The model system was built based on the crystallographic structure of Prp2 in complex with an ATP mimic (ADP-BeF$_3^-$, PDB ID: 6ZM2).[44] The system was solvated with a ~114 Å$^3$ cubic box of water molecules (TIP3P)[8] with a minimum distance of 15 Å between any protein atom and the edge of the box. The water molecules from the crystal structure were preserved. Finally, a total of 134 Na$^+$ ions and 121 Cl$^-$ ions were added to neutralize and to set an ion concentration of 150 mM. The number of ions was computed with the server and protocols provided by Schmit and colleagues.[48] After an initial equilibration of 150 ns-long MD simulations, the system was further relaxed for 10 ps by a QM/MM MD simulation, with the BLYP[49,50] functional describing the QM region.

We performed QM/MM metadynamics simulations to compute the free energy profile along the ATP hydrolysis mechanism. This allows us to unlock the ATPase mechanism and see how this is activated by RNA binding. To this end, we carried out metadynamics simulations to explore the whole free energy profile in the presence of RNA. In addition, metadynamics simulations of the rate-determining step were also done in the absence of RNA. In the last case, the large flexibility of the general base makes any recrossing event extremely unlikely. For this reason, in the absence of the RNA strand, the free energy barrier of the rate-determining step was calculated by performing multiple metadynamics runs and by stopping the simulation after the transition state has been overcome. This approach has been successfully used in other computational studies.[51]

In the presence of the RNA, at each step of the catalytic mechanism, the height of the added hills was set to 0.6125 $\text{kcal} \cdot \text{mol}^{-1}$ (<kT at 310 K) and the width of the Gaussian functions was set up according with the oscillations of the collective variables (CVs) in unbiased QM/MM MD simulations.[52] Gaussians functions were added every 30 fs corresponding to 120 or 60 MD steps (depending on the time step used). For chemical steps we used a time step of 0.25 fs while for the proton

networks rearrangements time step was set to 0.5 fs. As proposed by Ensing et. al., the QM/MM metadynamics were stopped after all the critical points (principal minima and transition states) were sampled at least once.[52] In each case, three replicas were carried out starting with half of the first minimum already filled. The height of the hills for those replicas was reduced to 3/4, 1/2 and 1/4 of the original value.

Since in the absence of RNA, the active site is more flexible, and the position of the general base is quite variable, six replicas of the rate-determining step were performed to assess the reproducibility of the results. A first complete filling was done using 0.6125 $\text{kcal} \cdot \text{mol}^{-1}$ as the height of hill. Then, 5 extra replicas were done, starting with half of the first minimum already filled. The height of the hills for those replicas was 3/4, 1/2, 1/4, 1/6 and 1/8 of the original value.

In all cases, variances in the barriers, along the performed replicas using different hill heights, were used to calculate the error following the protocol proposed by Nair et al.[53] The molecular mechanism resulted the same for all the Gaussian heights used. It is well known that metadynamics proceeds via the minimum free energy path. Therefore, by selecting proper collective variables, we can directly discriminate between different reaction pathways and disregard the less favorable ones.

The equilibrated Michaelis complex (R, Figure 4.17) exhibited a stable structure similar to the crystallographic one (average RMSD of the active site = 1.65 Å) and maintained the same hydrogen bond network surrounding the ATP substrate. Namely, Arg625 interacts with ATP γ-phosphate, while Lys326 and Arg628 interact with both the γ and β phosphates (Figure 4.17). Arg625, besides interacting with the γ phosphate of ATP, also interacts with the Oε:Gln621. The Glu419, Ser578 and Gln621, in turn, H-bond with the catalytic water, hence contributing to retain the nucleophilic oxygen ($O_{nuc}$:Wcat) at the optimal distance (3.3±0.2 Å) from the Pγ:ATP and in line orientation (161±0.2°) for the nucleophilic attack. Consistently with the crystal structure, a second water molecule (W2) settles within the active site (Figure 4.17), acting as hydrogen bond donor to the Oγ:ATP and the Oε:Glu419, while acting as hydrogen bond acceptor of the HNε:Gln621 and the HN:Ala451.

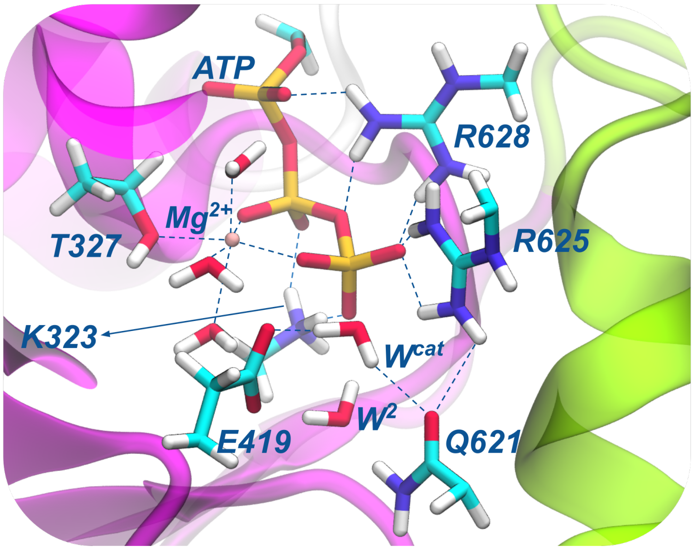


***Figure 4.17.*** *General disposition of the Prp2 active site as reactant state in the ATP hydrolysis reaction. The H-bond interactions are shown as dashed lines. The region treated quantum mechanically to explore the mechanism of ATP hydrolysis is shown in licorice representation with hydrogen atoms explicitly. The W2 molecule was included in the quantum region after the first step (**R** → **I1**) of the hydrolysis.*

Next, we inspected the first step of the ATP hydrolysis reaction (R→I1, Figure 4.18 and Figure 4.19) by promoting the nucleophilic attack of the $O_{nuc}$:Wcat on the Pγ:ATP in QM/MM MTD simulation. To this end, we used as collective variable (CV1) the difference between the distances of the forming bond (i.e. distance between Onuc:Wcat and Pγ:ATP) and the breaking bond (i.e. the distance between Oβ:ATP and Pγ:ATP). The hydrolysis occurs through a synchronous concerted path where the attack of the nucleophilic water on the scissile phosphate in sync with the removal of one proton from Wcat by Glu419 residue, which acts as the general base of the catalytic process. The use of CV1 allows us to discard competitive pathways that pass through the formation of intermediates, such as metaphosphates or pentacoordinate phosphates. Moreover, even in the absence of a CV specifically biasing the proton transfer, the deprotonation of the nucleophilic water is always mediated by Glu419 in the four replicas of the first reaction step we performed. Thus, the assistance of any other nearby bases is unlikely.

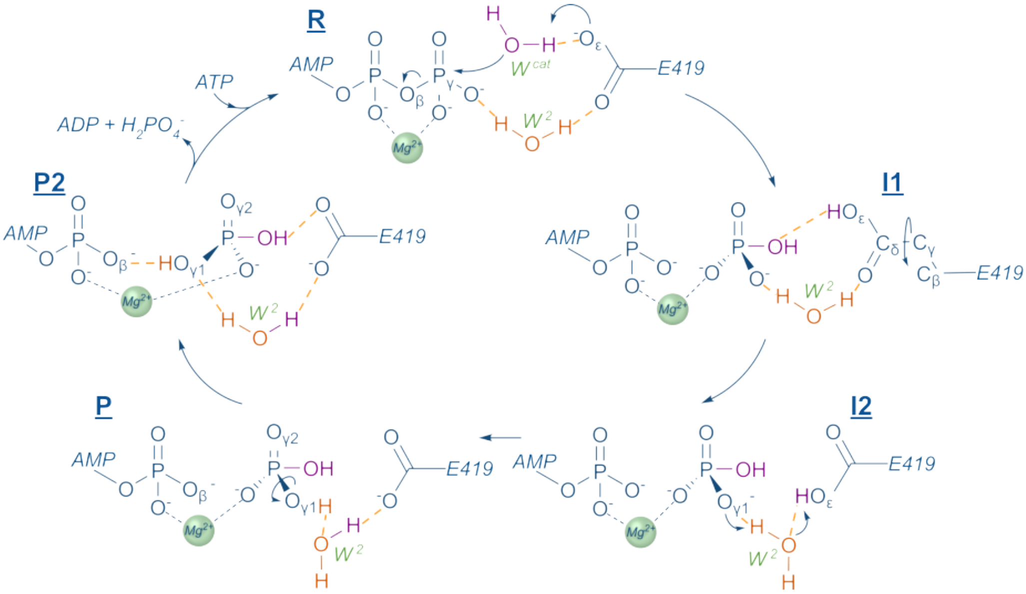


***Figure 4.18.*** *Sketch of the ATP hydrolysis reaction mechanism catalyzed by the DEAH-box Prp2 ATPase/helicase computed by metadynamics simulations at BLYP-D3/MM level.*

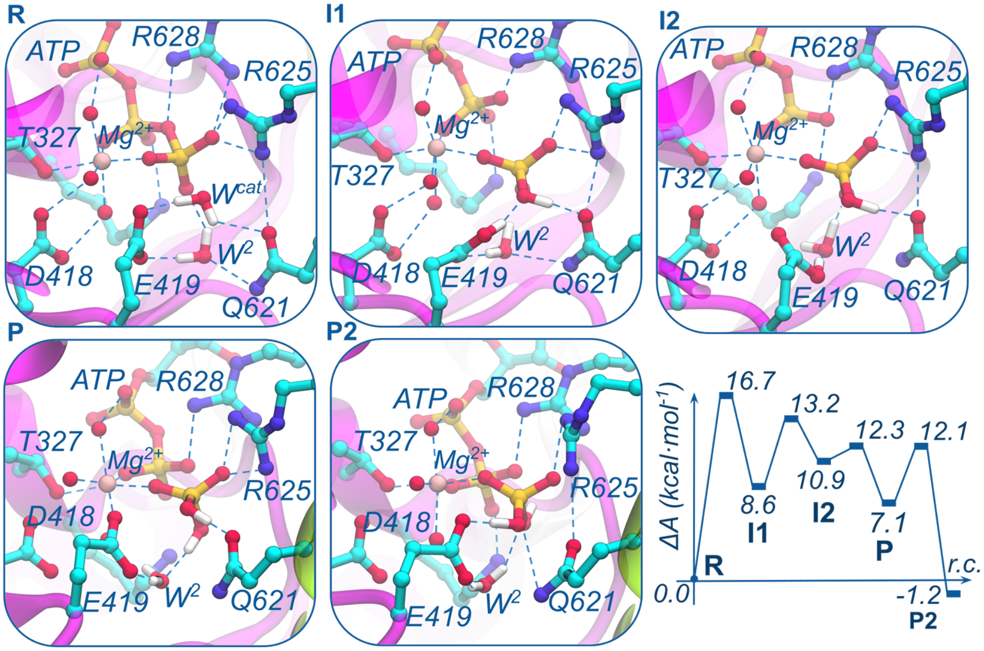


***Figure 4.19.*** *Representative structures of the key states of the ATP hydrolysis mechanism catalyzed by Prp2 and activated by the binding of an RNA strand. Lower right panel: Schematic representation of the free energy profile of the ATP hydrolysis mechanism calculated at QM(BLYP-D3(DZVP))/MM level. Reaction free energies and free energy barriers are reported with respect to the reactant state in kcal · mol$^{-1}$.*

First step, leading to the formation of a $HPO_4^{2-}$ and an ADP molecule, occurs by overcoming a Helmholtz free energy barrier ($\Delta A^{\ddagger}$) of 16.7±0.8 kcal · mol$^{-1}$ (Figure 4.19). Close to the transition state ($TS^{R \rightarrow I1}$), the distances of the forming and breaking bonds (Onuc:Wcat-Pγ:ATP and Pγ:ATP-Oβ:ATP) are both close to 2.2 Å (Figure 4.20). Notably, the calculated $\Delta A^{\ddagger}$ is in close agreement with the experimentally measured kcat of 3.2 s$^{-1}$, which by applying the Eyring equation,[54] in the context of the Transition State Theory (TST),[55] corresponding to a Gibbs free energy barrier ($\Delta G^{\ddagger}$) of 17.4 kcal · mol$^{-1}$ at 310 K.

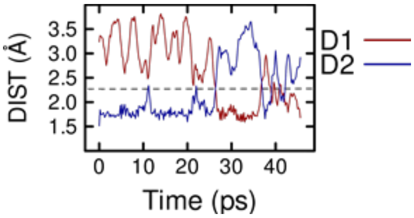


***Figure 4.20.*** *Time evolution of D1 ($O_{nuc}$:Wcat -Pγ:ATP) and D2 (Pγ:ATP- Oβ:ATP) distances along the MTD simulation (ps) of the rate determinant R→I1 step in the presence of RNA. The approximated transition state distance is shown with a dashed line.*

The catalytic reaction further proceeds (I1→I2) with a rearrangement of the hydrogen bond network of the Glu419 carboxylic group. After undergoing a rotation of the Cβ-Cγ-Cδ-OHε:Glu419 dihedral angle by ~35°, HOε:Glu419 is oriented towards the oxygen of a second water molecule (W2, Figure 4.18 and Figure 4.19) located in the active site. As a result, W2 also rearranges to hydrogen bond a negatively charged oxygen atom of the newly formed $HPO_4^{2-}$ ion. In this manner, W2 bridges the $HPO_4^{2-}$ ion and HOε:Glu419. We estimate that the Glu419 conformational rearrangement occurs at a $\Delta A^‡$ of 4.6±0.7 kcal · mol$^{-1}$ ($TS^{I1\rightarrow I2}$). This value was estimated by performing a MTD simulation, using as CV2 the difference of distances of the hydrogen of the general base HOε:Glu419 and the nucleophilic $O_{nuc}$:Wcat, which, after the first catalytic step, is part of the $HPO_4^{2-}$ ion, minus the distance of HOε:Glu419 and O:W2. As a result, a metastable intermediate (I2, Figures 4.18 and 4.19) is formed ($\Delta A$=2.3 kcal · mol$^{-1}$ from I1), which is rapidly converted into products (P) through a $\Delta A^‡$ of 1.4 kcal · mol$^{-1}$ ($TS^{I2\rightarrow P}$).

The formation of products occurs via an asynchronous double proton transfer first from H:W2 to Oγ1:$HPO_4^{2-}$ and then from HOε:Glu419 to O:W2. This process is achieved by performing a MTD simulation with two CVs (i.e. CV3 equal to the difference of the distances between HOε:Glu419 and Oε:Glu419 and the distance between HOε:Glu419 and O:W2, and CV3' equal to the difference of distances between H:W2 and O:W2 and the distance between H:W2 and Oγ1:$HPO_4^{2-}$). The almost spontaneous character of this process is due to the basicity of the $HPO_4^{2-}$ ion at physiological conditions. In addition to this pathway, we tried to evaluated the proton transfer directly to the other oxygen of $HPO_4^{2-}$ ion (the one coordinated to $Mg^{2+}$). However, because its participation in the coordination sphere decreases its basicity, this alternative path resulted in very high energies and distortion of the $Mg^{2+}$ ion coordination sphere. Furthermore, the crystal structures with ATP[44] and ADP[56,57] suggest that i) the second water molecule may participates in the ATP hydrolisis reaction and ii) the coordination sphere remains invariant after hydrolysis, thus supporting our findings.

P still lies at an $\Delta A$ of +7.1 kcal · $mol^{-1}$ as compared to reactant R. The formation of an energetically favorable product (P2, Figures 4.18 and 4.19), occurs only after an hydrogen bond network rearrangement. Here the proton of the $HO\gamma1{:}H_2PO_4^-$ reorients towards the $O\beta$:ADP. A similar mechanism of product stabilizations was reported also for other ATPases, and may be instrumental to reduce the interaction of $H_2PO_4^-$ with the catalytic site residues, thus facilitating its release. This rearrangement occurs by overcoming a $\Delta A^\ddagger$ of 5.0±1.3 kcal · $mol^{-1}$ and is achieved by performing a MTD simulation using as CV4 the dihedral angle $HO\gamma1$-$O\gamma1$ -$P\gamma$-$O\gamma2{:}H_2PO_4^-$, (Figures 4.18 and 4.19). This step leads to an overall exergonic catalytic process ($\Delta A$ = -1.2 kcal · $mol^{-1}$).

It is well known that Prp2 as well as all helicases of the DExH-box family are RNA-dependent ATPases.[41] In order to explore the role of the RNA strand in activating the Prp2 ATPase function we equilibrated the system with 1 μs-long MD simulation in the absence of the RNA and we performed 6 MTD replicas of the nucleophilic attack (i.e. the rate determining step). As a result, the calculated $\Delta A^\ddagger$ raises to 23.5±3.2 kcal · $mol^{-1}$ (Figure 4.21). The large standard deviation of this free energy barrier, by comparison with the value obtained in the presence of RNA, is due to the increased flexibility of the general base Glu419 (*vide infra*).

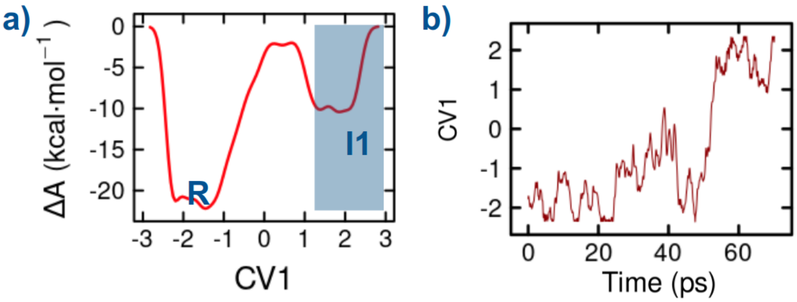


***Figure 4.21.*** *a) Helmholtz free energy ($\Delta A$) profile vs. CV1 = [Onuc :Wcat-Pγ:ATP]-[Pγ:ATP- Oβ:ATP] for the system without RNA. $\Delta A$ was computed at DFT-BLYP-D3/MM level for ATP hydrolysis catalyzed by Prp2 and is reported in kcal·mol$^{-1}$. b) Time evolution of the collective variable CV1 along the MTD trajectories (ps) for the rate determinant step in the absence of RNA.*

**The RNA client allosterically places the catalytic water molecule in a reactive configuration.** In order to inspect the way the single strand RNA allosterically triggers the activation of Prp2 ATPase/helicase function, we carried out four independent 1 µs-long classical MD simulations on the reactant state in the presence and absence of a seven nucleotides-long polyU RNA strand. Simulations parameters were the same for the previous studies but the simulations run at NPT ensemble, 310 K and 1 bar. Surprisingly, the structure of the active site is very similar in the two cases. Nonetheless, the RNA strand constraints the reactive disposition of Glu419 and Ser578 residues lining the active site. Indeed, the flexibility (root mean square fluctuation in Figure 4.22 and Table 4.2) of these residues markedly raises in the absence of the RNA strand.

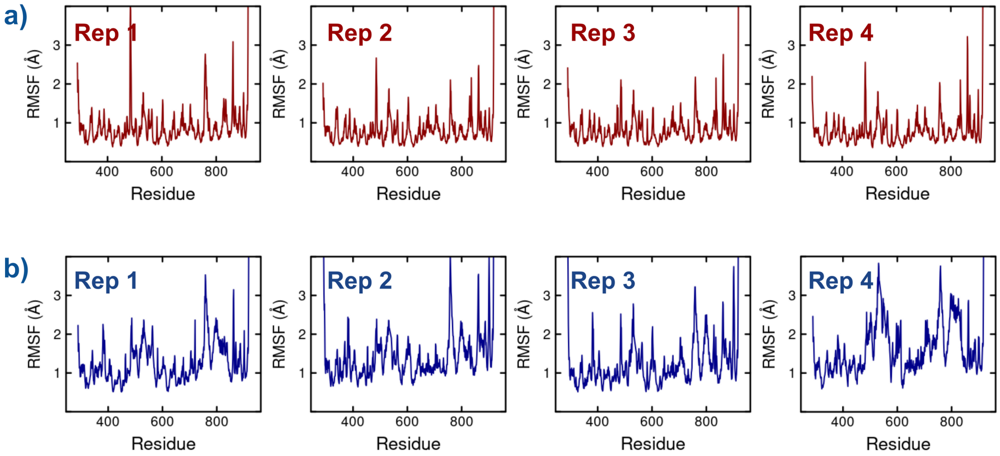


***Figure 4.22.*** *Root-mean-square fluctuation (RMSF) analysis of protein Cα atoms using the initial reactive structure as reference for the four replicas of the classical molecular dynamics simulations a) in presence and b) in absence of RNA.*

***Table 4.2.*** *Average values of the root-mean-square fluctuation (RMSF) of Cα atoms using the initial reactive structure as reference for the residues involved in the RNA-driven allosteric activation of the Prp2 ATPase/helicase function along the four classical MD replicas.*

| **Resid** | **RMSF wRNA (Å)** | **RMSF woRNA (Å)** |
|---|---|---|
| Arg352 | 0.51±0.07 | 1.10±0.26 |
| Arg353 | 0.48±0.04 | 1.16±0.21 |
| Glu419 | 0.39±0.01 | 0.67±0.18 |
| Glu422 | 0.46±0.04 | 0.87±0.23 |
| Asn573 | 0.47±0.03 | 1.08±0.29 |
| Ser578 | 0.44±0.03 | 0.71±0.16 |

As detailed above, Glu419, besides taking direct part in most steps of the enzymatic mechanism, also properly orients the catalytic water Wcat for the nucleophilic attack, which is the rate determining step of the ATP hydrolysis. The increased flexibility of Glu419 and Ser578 remodels the hydrophilic cavity in the vicinity of P$\gamma$:ATP, increasing its volume. This enables the entrance of an extra water molecule into the active site (W3, Figure 4.23) and, most importantly, triggers a repositioning of the nucleophilic Wcat. Briefly, in the lack of the RNA strand, the Wcat no longer explores the optimal and indispensable in-line position for the nucleophilic attack (Figure 4.23).

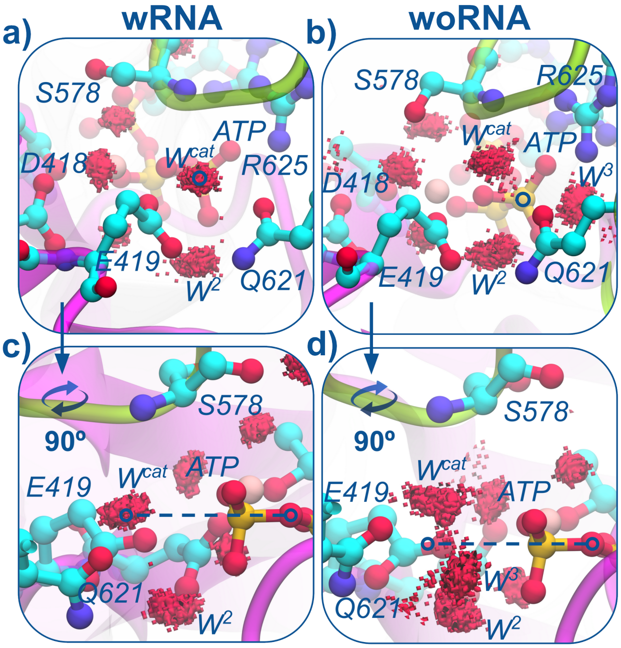


***Figure 4.23.*** *Representative frames of the simulations of Prp2 in the presence (left) and absence (right) of the RNA strand. Front view of the density water density distribution in the active site in a) presence and b) absence of the RNA strand. Side view of the density water density distribution in the active site in a) presence and b) absence of the RNA strand. The ideal position of the nucleophilic water is depicted by a blue circle and the ideal alignment for the attack to the P$\gamma$:ATP atom is highlighted by blue dashed lines.*

In order to capture the molecular signature underlying this RNA-mediate allosteric activation of Prp2 catalysis, we closely inspected the residues mechanically coupling the catalytic site and the RNA binding tunnel. As a result, we observed that in the presence of the RNA strand Arg353 engages persistent hydrogen bond and salt-bridge interactions (with its NH backbone and guanidinium moieties) to the phosphate backbone of U5 and U6 nucleotides, respectively. This in turn rigidifies the flanking Arg352 residue whose guanidinium moiety engages a salt-bridge to the side chain of Glu422 (Figure 4.24). In Figure 4.24 is depicted the distance distribution between Cδ:Glu419-Cζ:Arg352 as D1 showing larger values in the absence of RNA (4.69±0.5 and 6.25±1.22 Å in the presence/absence of RNA, respectively, Figure 4.24). Moreover, Glu422 and the Glu419 belong to the same loop and directly hydrogen bonds through their backbone atoms. Then, the salt-bridge between Arg352 and Glu422 locks the position of the Glu419 in the presence of RNA, the general base of the ATPase reaction, into the optimal position to activate the nucleophilic water Wcat (Figure 4.24). While in the absence of RNA, Glu422 and Glu419 are more flexible and the latter does not establish a proper position to preorient and to activate the nucleophilic water Wcat to attack Pγ:ATP. This provides an explanation to the markedly larger free energy barrier calculated in the absence of the RNA (Figure 4.21).

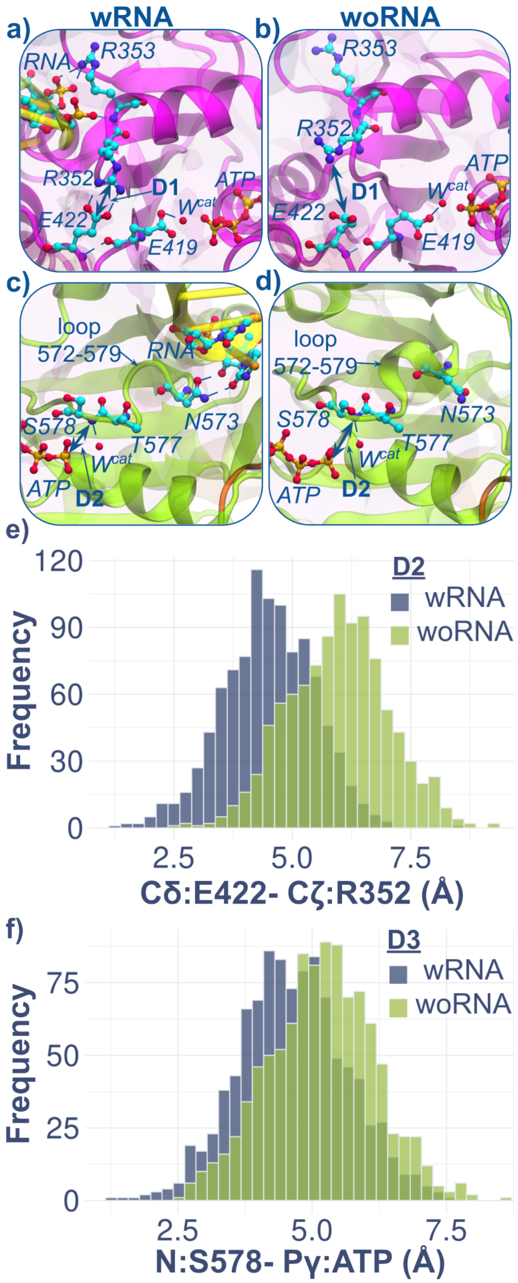


***Figure 4.24.*** *Representative frames with the orientation of the key Prp2 residues, ATP substrate and catalytic water in the presence (panels a and c) and absence (panels b and d) the RNA strand. e) Distribution of the distance (Å) between Cδ:Glu419-*

*Cζ:Arg352 atoms (marked as D1 in a) and b)) along the 1 μs long classical MD trajectory in the presence (blue) and absence (green) of the RNA strand. f) Distribution of the distance (in Å) between N:Ser578-Pγ:ATP atoms (marked as D2 in c) and d)) along the 1-μs long classical MD in the presence (blue) and absence (green) of the RNA strand.*

Additionally, we identified a second set of important interactions. Namely, the side chain of Asn573 interacts with the sugar of U4 nucleotide, rigidifying the loop formed by residues 572 - 579 where Ser578 belongs to (Figure 4.24). In the absence of the RNA strand HN:Ser578, which faces the active site, becomes more flexible (Figure 4.22 and Table 4.2) and lies at larger distances from Pγ:ATP (4.78±0.2 vs 5.07±0.2 Å in the presence/absence of RNA, respectively, D2 in Figure 4.24). This shift in the relative position of Ser578 strongly contributes to the Wcat repositioning, thus leading to impaired catalysis. Remarkably, Asn573 and Arg353 interact with RNA through one sugar and two phosphates, respectively, which is congruent with the sequence-independent nature of Prp2. The observed trends were confirmed in all MD simulations replicas.

A bioinformatic analysis done with the Protein BLAST server strikingly revealed that the residues involved in the allosteric activation of Prp2 ATP hydrolysis catalysis are highly conserved among other DExH-box ATPase/helicases (Figure 4.25). Namely, the Arg352, Arg353, Glu419, Glu422 (hereafter name RR-EE motif), and the Asn573 and Ser578 (NS motif) residues are invariant among the functional orthologs of Prp2 and in the other spliceosomal DExH-box ATPase/helicases Prp16, Prp22, Prp43. The high degree of conservation suggests that any mutation (natural or *in vitro*) of the aminoacids involved either in allosteric communication from the RNA binding site to the catalytic site (Arg352, Arg353, Glu419, Glu422 and Asn573 and Ser578) or in the activation of the nucleophilic water (Glu419) directly impacts the ATPase/helicase activity of any member of the DExH-box family.

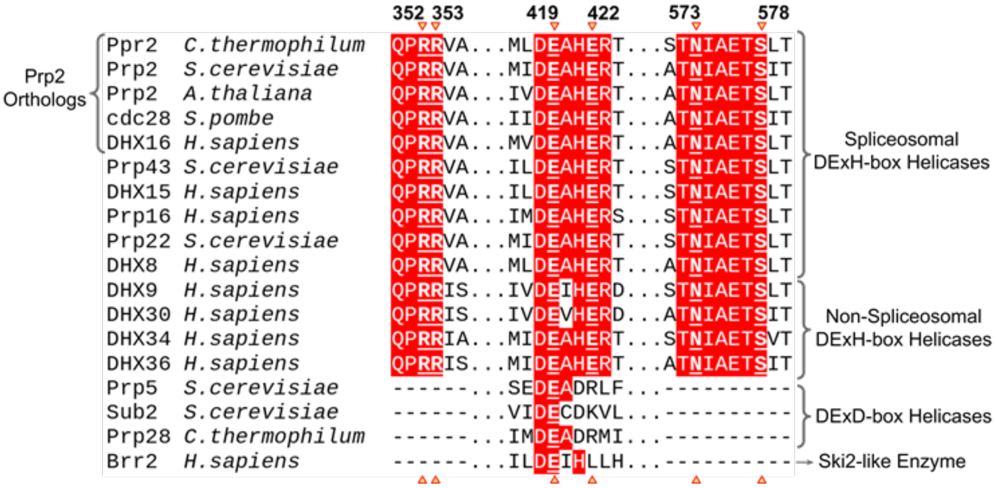


**Figure 4.25.** *Sequence alignment among Prp2, various DExH-box ATPases/helicases, spliceosomal DExD-box ATPases/helicases and Brr2 helicase of the Ski-like family. The conserved residues are shaded red and those involved in allosteric modulation are marked in bold and underlined.*

The human genome encodes 16 DExH-box RNA helicases that are involved in a variety of key biological processes such as homologous recombination repair,[58] nonsense decay activation mediation,[59] unwinding of G4 structures forming in the 3′-untranslated regions of mRNA,[60] ribosome biogenesis, global translation, and mitochondrial metabolism.[61] Stunningly, the residues involved in allosteric activation of Prp2 ATPase function are conserved even among several other DExH-box ATPase/helicases not taking part to the splicing cycle (Figure 4.25), suggesting that this allosteric regulation may be broadly applicable to all DExH helicases. This mechanism is nonetheless unique to the DExH subfamily as revealed by the sequence alignment of other spliceosomal helicases from the DExD-box and Ski2-like families (Figure 4.25), where instead these key motifs are lacking. More broadly, this mechanism resembles the so called ‘glutamate switch’ underlying the allosteric activation in AAA+ family of proteins.[62] In AAA+ enzymes, a positively charged Arg/Lys or polar Gln/Asn/Thr/Ser act as Glu-switch residues are proposed to thwart ATP hydrolysis by hydrogen bonding to a catalytic Glu residues and holding it in a catalytically inactive conformation until the external signal (nucleic acid and substrate ATP binding) occurs. In DExH-box helicases, instead, RNA binding/dissociation switches on/off the salt-bridge between Arg and Glu residues, in the center of a ‘RR-EE’ motif where the first residue directly interacts with RNA and the latter is the catalytic Glu419 that acts as base. The formation of this salt-bridges contributes to mechanically regulating the RNA-driven activation of this pivotal protein family.

The good agreement between structural and kinetic available experimental data and the results derived from our simulations confirms that MTD can be an efficient method to explore not only conformational changes but free energy surfaces of chemical reactions of complex biological system, such as the ATP hydrolysis catalyzed by Prp2 and activated by the binding of an RNA strand, at DFT/MM level. Consequently this method can be considered as a promising alternative to study the inhibition process of RgpB, where the presence of a sulphur atom in the reactive system can be a possible souce of error when employing semiempirical methods to treat the QM region.

## 4.4 REFERENCES

(1) McDonald, I. R.; Singer, K. Machine Calculation of Thermodynamic Properties of a Simple Fluid at Supercritical Temperatures. *J. Chem. Phys.* **1967**, *47* (11), 4766–4772. https://doi.org/10.1063/1.1701695.

(2) Torrie, G. M.; Valleau, J. P. Nonphysical Sampling Distributions in Monte Carlo Free-Energy Estimation: Umbrella Sampling. *J. Comput. Phys.* **1977**, *23* (2), 187–199. https://doi.org/10.1016/0021-9991(77)90121-8.

(3) Dominy, S. S.; Lynch, C.; Ermini, F.; Benedyk, M.; Marczyk, A.; Konradi, A.; Nguyen, M.; Haditsch, U.; Raha, D.; Griffin, C.; Holsinger, L. J.; Arastu-Kapur, S.; Kaba, S.; Lee, A.; Ryder, M. I.; Potempa, B.; Mydel, P.; Hellvard, A.; Adamowicz, K.; Hasturk, H.; Walker, G. D.; Reynolds, E. C.; Faull, R. L. M.; Curtis, M. A.; Dragunow, M.; Potempa, J. Porphyromonas Gingivalis in Alzheimer's Disease Brains: Evidence for Disease Causation and Treatment with Small-Molecule Inhibitors. *Sci. Adv.* **2019**, *5* (1), eaau3333. https://doi.org/10.1126/sciadv.aau3333.

(4) Bateman, A.; Martin, M.-J.; Orchard, S.; Magrane, M.; Agivetova, R.; Ahmad, S.; Alpi, E.; Bowler-Barnett, E. H.; Britto, R.; Bursteinas, B.; Bye-A-Jee, H.; Coetzee, R.; Cukura, A.; Da Silva, A.; Denny, P.; Dogan, T.; Ebenezer, T.; Fan, J.; Castro, L. G.; Garmiri, P.; Georghiou, G.; Gonzales, L.; Hatton-Ellis, E.; Hussein, A.; Ignatchenko, A.; Insana, G.; Ishtiaq, R.; Jokinen, P.; Joshi, V.; Jyothi, D.; Lock, A.; Lopez, R.; Luciani, A.; Luo, J.; Lussi, Y.; MacDougall, A.; Madeira, F.; Mahmoudy, M.; Menchi, M.; Mishra, A.; Moulang, K.; Nightingale, A.; Oliveira, C. S.; Pundir, S.; Qi, G.; Raj, S.; Rice, D.; Lopez, M. R.; Saidi, R.; Sampson, J.; Sawford, T.; Speretta, E.; Turner, E.; Tyagi, N.; Vasudev, P.; Volynkin, V.; Warner, K.; Watkins, X.; Zaru, R.; Zellner, H.; Bridge, A.; Poux, S.; Redaschi, N.; Aimo, L.; Argoud-Puy, G.; Auchincloss, A.; Axelsen, K.; Bansal, P.; Baratin, D.; Blatter, M.-C.; Bolleman, J.; Boutet, E.; Breuza, L.; Casals-Casas, C.; de Castro, E.; Echioukh, K. C.; Coudert, E.; Cuche, B.; Doche, M.; Dornevil, D.; Estreicher, A.; Famiglietti, M. L.; Feuermann, M.; Gasteiger, E.; Gehant, S.; Gerritsen, V.; Gos, A.; Gruaz-Gumowski, N.; Hinz, U.; Hulo, C.; Hyka-Nouspikel, N.; Jungo, F.; Keller, G.; Kerhornou, A.; Lara, V.; Le Mercier, P.; Lieberherr, D.; Lombardot, T.; Martin, X.; Masson, P.; Morgat, A.; Neto, T.

B.; Paesano, S.; Pedruzzi, I.; Pilbout, S.; Pourcel, L.; Pozzato, M.; Pruess, M.; Rivoire, C.; Sigrist, C.; Sonesson, K.; Stutz, A.; Sundaram, S.; Tognolli, M.; Verbregue, L.; Wu, C. H.; Arighi, C. N.; Arminski, L.; Chen, C.; Chen, Y.; Garavelli, J. S.; Huang, H.; Laiho, K.; McGarvey, P.; Natale, D. A.; Ross, K.; Vinayaka, C. R.; Wang, Q.; Wang, Y.; Yeh, L.-S.; Zhang, J.; Ruch, P.; Teodoro, D. UniProt: The Universal Protein Knowledgebase in 2021. *Nucleic Acids Res.* **2021**, *49* (D1), D480–D489. https://doi.org/10.1093/nar/gkaa1100.

(5) Potempa, J.; Mikolajczyk-Pawlinska, J.; Brassell, D.; Nelson, D.; Thøgersen, I. B.; Enghild, J. J.; Travis, J. Comparative Properties of Two Cysteine Proteinases (Gingipains R), the Products of Two Related but Individual Genes of Porphyromonas Gingivalis. *J. Biol. Chem.* **1998**, *273* (34), 21648–21657. https://doi.org/10.1074/jbc.273.34.21648.

(6) Maier, J. A.; Martinez, C.; Kasavajhala, K.; Wickstrom, L.; Hauser, K. E.; Simmerling, C. Ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from Ff99SB. *J. Chem. Theory Comput.* **2015**, *11* (8), 3696–3713. https://doi.org/10.1021/acs.jctc.5b00255.

(7) Joung, I. S.; Cheatham, T. E. Determination of Alkali and Halide Monovalent Ion Parameters for Use in Explicitly Solvated Biomolecular Simulations. *J. Phys. Chem. B* **2008**, *112* (30), 9020–9041. https://doi.org/10.1021/jp8001614.

(8) 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.* **1983**, *79* (2), 926–935. https://doi.org/10.1063/1.445869.

(9) Essmann, U.; Perera, L.; Berkowitz, M. L.; Darden, T.; Lee, H.; Pedersen, L. G. A Smooth Particle Mesh Ewald Method. *J. Chem. Phys.* **1995**, *103* (19), 8577–8593. https://doi.org/10.1063/1.470117.

(10) Darden, T.; York, D.; Pedersen, L. Particle Mesh Ewald: An N ⋅log( N ) Method for Ewald Sums in Large Systems. *J. Chem. Phys.* **1993**, *98* (12), 10089–10092. https://doi.org/10.1063/1.464397.

(11) Uberuaga, B. P.; Anghel, M.; Voter, A. F. Synchronization of Trajectories in Canonical Molecular-Dynamics Simulations: Observation, Explanation, and

Exploitation. *J. Chem. Phys.* **2004**, *120* (14), 6363–6374. https://doi.org/10.1063/1.1667473.

(12) Sindhikara, D. J.; Kim, S.; Voter, A. F.; Roitberg, A. E. Bad Seeds Sprout Perilous Dynamics: Stochastic Thermostat Induced Trajectory Synchronization in Biomolecules. *J. Chem. Theory Comput.* **2009**, *5* (6), 1624–1631. https://doi.org/10.1021/ct800573m.

(13) Ryckaert, J.-P.; Ciccotti, G.; Berendsen, H. J. Numerical Integration of the Cartesian Equations of Motion of a System with Constraints: Molecular Dynamics of n-Alkanes. *J. Comput. Phys.* **1977**, *23* (3), 327–341. https://doi.org/10.1016/0021-9991(77)90098-5.

(14) Kräutler, V.; van Gunsteren, W. F.; Hünenberger, P. H. A Fast SHAKE Algorithm to Solve Distance Constraint Equations for Small Molecules in Molecular Dynamics Simulations. *J. Comput. Chem.* **2001**, *22* (5), 501–508. https://doi.org/10.1002/1096-987X(20010415)22:5<501::AID-JCC1021>3.0.CO;2-V.

(15) Verlet, L. Computer "Experiments" on Classical Fluids. I. Thermodynamical Properties of Lennard-Jones Molecules. *Phys. Rev.* **1967**, *159* (1), 98–103. https://doi.org/10.1103/PhysRev.159.98.

(16) Movilla, S.; Martí, S.; Roca, M.; Moliner, V. Unrevealing the Proteolytic Activity of RgpB Gingipain from Computational Simulations. *J. Chem. Inf. Model.* **2021**, *61* (9), 4582–4593. https://doi.org/10.1021/acs.jcim.1c00666.

(17) Eichinger, A.; Beisel, H. G.; Jacob, U.; Huber, R.; Medrano, F. J.; Banbula, A.; Potempa, J.; Travis, J.; Bode, W. Crystal Structure of Gingipain R: An Arg-Specific Bacterial Cysteine Proteinase with a Caspase-like Fold. *EMBO J.* **1999**, *18*, 5453–5462. https://doi.org/10.1093/emboj/18.20.5453.

(18) Mazmanian, K.; Sargsyan, K.; Grauffel, C.; Dudev, T.; Lim, C. Preferred Hydrogen-Bonding Partners of Cysteine: Implications for Regulating Cys Functions. *J. Phys. Chem. B* **2016**, *120* (39), 10288–10296. https://doi.org/10.1021/acs.jpcb.6b08109.

(19) Elsässer, B.; Zauner, F. B.; Messner, J.; Soh, W. T.; Dall, E.; Brandstetter, H. 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.* **2017**, *7* (9), 5585–5593. https://doi.org/10.1021/acscatal.7b01505.

(20) Rangarajan, M.; Smith, S. J. M.; U, S.; Curtis, M. A. Biochemical Characterization of the Arginine-Specific Proteases of Porphyromonas Gingivalis W50 Suggests a Common Precursor. *Biochem. J.* **1997**, *323* (3), 701–709. https://doi.org/10.1042/bj3230701.

(21) Ally, N.; Whisstock, J. C.; Sleprawska-Lupa, M.; Potempa, J.; Le Bonniec, B. F.; Travis, J.; Pike, R. N. Characterization of the Specificity of Arginine-Specific Gingipains from Porphyromonas Gingivalis Reveals Active Site Differences between Different Forms of the Enzymes. *Biochemistry* **2003**, *42* (40), 11693–11700. https://doi.org/10.1021/bi0349726.

(22) Hanada, K.; Tamai, M.; Yamagishi, M.; Ohmura, S.; Sawada, J.; Tanaka, I. Isolation and Characterization of E–64, a New Thiol Protease Inhibitor. *Agric. Biol. Chem.* **1978**, *42* (3), 523–528. https://doi.org/10.1080/00021369.1978.10863014.

(23) Arafet, K.; Ferrer, S.; Moliner, V. Computational Study of the Catalytic Mechanism of the Cruzain Cysteine Protease. *ACS Catal.* **2017**, *7* (2), 1207–1215. https://doi.org/10.1021/acscatal.6b03096.

(24) Świderek, K.; Moliner, V. Revealing the Molecular Mechanisms of Proteolysis of SARS-CoV-2 Mproby QM/MM Computational Methods. *Chem. Sci.* **2020**, *11* (39), 10626–10630. https://doi.org/10.1039/d0sc02823a.

(25) Dominy, S. S.; Lynch, C.; Ermini, F.; Benedyk, M.; Marczyk, A.; Konradi, A.; Nguyen, M.; Haditsch, U.; Raha, D.; Griffin, C.; Holsinger, L. J.; Arastu-Kapur, S.; Kaba, S.; Lee, A.; Ryder, M. I.; Potempa, B.; Mydel, P.; Hellvard, A.; Adamowicz, K.; Hasturk, H.; Walker, G. D.; Reynolds, E. C.; Faull, R. L. M.; Curtis, M. A.; Dragunow, M.; Potempa, J. Porphyromonas Gingivalis in Alzheimer's Disease Brains: Evidence for Disease Causation and Treatment with

Small-Molecule Inhibitors. *Sci. Adv.* **2019**, *5* (1), eaau3333. https://doi.org/10.1126/sciadv.aau3333.

(26) Duan, L.; Liu, X.; Zhang, J. Z. H. Interaction Entropy: A New Paradigm for Highly Efficient and Reliable Computation of Protein–Ligand Binding Free Energy. *J. Am. Chem. Soc.* **2016**, *138* (17), 5722–5728. https://doi.org/10.1021/jacs.6b02682.

(27) Hou, T.; Wang, J.; Li, Y.; Wang, W. Assessing the Performance of the MM/PBSA and MM/GBSA Methods. 1. The Accuracy of Binding Free Energy Calculations Based on Molecular Dynamics Simulations. *J. Chem. Inf. Model.* **2011**, *51* (1), 69–82. https://doi.org/10.1021/ci100275a.

(28) Jorgensen, W. L.; Ravimohan, C. Monte Carlo Simulation of Differences in Free Energies of Hydration. *J. Chem. Phys.* **1985**, *83* (6), 3050–3054. https://doi.org/10.1063/1.449208.

(29) Loeffler, H. H.; Bosisio, S.; Duarte Ramos Matos, G.; Suh, D.; Roux, B.; Mobley, D. L.; Michel, J. Reproducibility of Free Energy Calculations across Different Molecular Simulation Software Packages. *J. Chem. Theory Comput.* **2018**, *14* (11), 5567–5582. https://doi.org/10.1021/acs.jctc.8b00544.

(30) Maragliano, L.; Fischer, A.; Vanden-Eijnden, E.; Ciccotti, G. String Method in Collective Variables: Minimum Free Energy Paths and Isocommittor Surfaces. *J. Chem. Phys.* **2006**, *125* (2), 024106. https://doi.org/10.1063/1.2212942.

(31) Maragliano, L.; Vanden-Eijnden, E. On-the-Fly String Method for Minimum Free Energy Paths Calculation. *Chem. Phys. Lett.* **2007**, *446* (1–3), 182–190. https://doi.org/10.1016/j.cplett.2007.08.017.

(32) Pan, A. C.; Sezer, D.; Roux, B. Finding Transition Pathways Using the String Method with Swarms of Trajectories. *J. Phys. Chem. B* **2008**, *112* (11), 3432–3440. https://doi.org/10.1021/jp0777059.

(33) Vanden-Eijnden, E.; Venturoli, M. Revisiting the Finite Temperature String Method for the Calculation of Reaction Tubes and Free Energies. *J. Chem. Phys.* **2009**, *130* (19), 194103. https://doi.org/10.1063/1.3130083.

(34) Johnson, M. E.; Hummer, G. Characterization of a Dynamic String Method for the Construction of Transition Pathways in Molecular Reactions. *J. Phys. Chem. B* **2012**, *116* (29), 8573–8583. https://doi.org/10.1021/jp212611k.

(35) Zinovjev, K.; Tuñón, I. Adaptive Finite Temperature String Method in Collective Variables. *J. Phys. Chem. A* **2017**, *121* (51), 9764–9772. https://doi.org/10.1021/acs.jpca.7b10842.

(36) Martí, S.; Arafet, K.; Lodola, A.; Mulholland, A. J.; Świderek, K.; Moliner, V. Impact of Warhead Modulations on the Covalent Inhibition of SARS-CoV-2 M pro Explored by QM/MM Simulations. *ACS Catal.* **2022**, *12* (1), 698–708. https://doi.org/10.1021/acscatal.1c04661.

(37) Arafet, K.; Serrano-Aparicio, N.; Lodola, A.; Mulholland, A. J.; González, F. V.; Świderek, K.; Moliner, V. Mechanism of Inhibition of SARS-CoV-2 M pro by N3 Peptidyl Michael Acceptor Explained by QM/MM Simulations and Design of New Derivatives with Tunable Chemical Reactivity. *Chem. Sci.* **2021**, *12* (4), 1433–1444. https://doi.org/10.1039/D0SC06195F.

(38) Arafet, K.; Ferrer, S.; González, F. V.; Moliner, V. Quantum Mechanics/Molecular Mechanics Studies of the Mechanism of Cysteine Protease Inhibition by Peptidyl-2,3-Epoxyketones. *Phys. Chem. Chem. Phys.* **2017**, *19* (20), 12740–12748. https://doi.org/10.1039/C7CP01726J.

(39) Laio, A.; Parrinello, M. Escaping Free-Energy Minima. *Proc. Natl. Acad. Sci.* **2002**, *99* (20), 12562–12566. https://doi.org/10.1073/pnas.202427399.

(40) Roy, J.; Kim, K.; Maddock, J. R.; Anthony, J. G.; Woolford, J. L. The Final Stages of Spliceosome Maturation Require Spp2p That Can Interact with the DEAH Box Protein Prp2p and Promote Step 1 of Splicing. *RNA* **1995**, *1* (4), 375–390.

(41) Warkocki, Z.; Schneider, C.; Mozaffari-Jovin, S.; Schmitzová, J.; Höbartner, C.; Fabrizio, P.; Lührmann, R. The G-Patch Protein Spp2 Couples the Spliceosome-Stimulated ATPase Activity of the DEAH-Box Protein Prp2 to Catalytic Activation of the Spliceosome. *Genes Dev.* **2015**, *29* (1), 94–107. https://doi.org/10.1101/gad.253070.114.

(42) Warkocki, Z.; Odenwälder, P.; Schmitzová, J.; Platzmann, F.; Stark, H.; Urlaub, H.; Ficner, R.; Fabrizio, P.; Lührmann, R. Reconstitution of Both Steps of Saccharomyces Cerevisiae Splicing with Purified Spliceosomal Components. *Nat. Struct. Mol. Biol.* **2009**, *16* (12), 1237–1243. https://doi.org/10.1038/nsmb.1729.

(43) Kim, S. H.; Lin, R. J. Spliceosome Activation by PRP2 ATPase Prior to the First Transesterification Reaction of Pre-MRNA Splicing. *Mol. Cell. Biol.* **1996**, *16* (12), 6810–6819. https://doi.org/10.1128/MCB.16.12.6810.

(44) Hamann, F.; Zimmerningkat, L. C.; Becker, R. A.; Garbers, T. B.; Neumann, P.; Hub, J. S.; Ficner, R. The Structure of Prp2 Bound to RNA and ADP-BeF$^{3-}$ Reveals Structural Features Important for RNA Unwinding by DEAH-Box ATPases. *Acta Crystallogr. Sect. D Struct. Biol.* **2021**, *77* (4), 496–509. https://doi.org/10.1107/S2059798321001194.

(45) Salomon-Ferrer, R.; Case, D. A.; Walker, R. C. An Overview of the Amber Biomolecular Simulation Package. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2013**, *3* (2), 198–210. https://doi.org/10.1002/wcms.1121.

(46) Li, Z.; Song, L. F.; Li, P.; Merz, K. M. Systematic Parametrization of Divalent Metal Ions for the OPC3, OPC, TIP3P-FB, and TIP4P-FB Water Models. *J. Chem. Theory Comput.* **2020**, *16* (7), 4429–4442. https://doi.org/10.1021/acs.jctc.0c00194.

(47) Li, P.; Merz, K. M. Taking into Account the Ion-Induced Dipole Interaction in the Nonbonded Model of Ions. *J. Chem. Theory Comput.* **2014**, *10* (1), 289–297. https://doi.org/10.1021/ct400751u.

(48) Schmit, J. D.; Kariyawasam, N. L.; Needham, V.; Smith, P. E. SLTCAP: A Simple Method for Calculating the Number of Ions Needed for MD Simulation. *J. Chem. Theory Comput.* **2018**, *14* (4), 1823–1827. https://doi.org/10.1021/acs.jctc.7b01254.

(49) 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* **1988**, *37* (2), 785–789. https://doi.org/10.1103/PhysRevB.37.785.

(50) Becke, A. D. Density-Functional Exchange-Energy Approximation with Correct Asymptotic Behavior. *Phys. Rev. A* **1988**, *38* (6), 3098–3100. https://doi.org/10.1103/PhysRevA.38.3098.

(51) Field, M. J.; Bash, P. A.; Karplus, M. A Combined Quantum Mechanical and Molecular Mechanical Potential for Molecular Dynamics Simulations. *J. Comput. Chem.* **1990**, *11* (6), 700–733. https://doi.org/10.1002/jcc.540110605.

(52) Davis, T. D.; Maggiora, G. M.; Christoffersen, R. E. Ab Initio Calculations on Large Molecules Using Molecular Fragments. Unrestricted Hartree-Fock Calculations of the Low-Lying States of Formaldehyde and Its Radical Ions. *J. Am. Chem. Soc.* **1974**, *96* (26), 7878–7887. https://doi.org/10.1021/ja00833a007.

(53) Laino, T.; Mohamed, F.; Laio, A.; Parrinello, M. An Efficient Real Space Multigrid QM/MM Electrostatic Coupling. *J. Chem. Theory Comput.* **2005**, *1* (6), 1176–1184. https://doi.org/10.1021/ct050123f.

(54) Eyring, H. The Activated Complex in Chemical Reactions. *J. Chem. Phys.* **1935**, *3* (2), 107–115. https://doi.org/10.1063/1.1749604.

(55) Truhlar, D. G.; Hase, W. L.; Hynes, J. T. Current Status of Transition-State Theory. *J. Phys. Chem.* **1983**, *87* (26), 5523–5523. https://doi.org/10.1021/j150644a044.

(56) Bai, R.; Wan, R.; Yan, C.; Jia, Q.; Lei, J.; Shi, Y. Mechanism of Spliceosome Remodeling by the ATPase/Helicase Prp2 and Its Coactivator Spp2. *Science (80-. ).* **2021**, *371* (6525). https://doi.org/10.1126/science.abe8863.

(57) Schmitt, A.; Hamann, F.; Neumann, P.; Ficner, R. Crystal Structure of the Spliceosomal DEAH-Box ATPase Prp2. *Acta Crystallogr. Sect. D Struct. Biol.* **2018**, *74* (7), 643–654. https://doi.org/10.1107/S2059798318006356.

(58) Chakraborty, P.; Hiom, K. DHX9-Dependent Recruitment of BRCA1 to RNA Promotes DNA End Resection in Homologous Recombination. *Nat. Commun.* **2021**, *12* (1), 4126. https://doi.org/10.1038/s41467-021-24341-z.

(59) Melero, R.; Hug, N.; López-Perrote, A.; Yamashita, A.; Cáceres, J. F.; Llorca, O. The RNA Helicase DHX34 Functions as a Scaffold for SMG1-Mediated UPF1 Phosphorylation. *Nat. Commun.* **2016**, *7* (1), 10585. https://doi.org/10.1038/ncomms10585.

(60) Sauer, M.; Juranek, S. A.; Marks, J.; De Magis, A.; Kazemier, H. G.; Hilbig, D.; Benhalevy, D.; Wang, X.; Hafner, M.; Paeschke, K. DHX36 Prevents the Accumulation of Translationally Inactive MRNAs with G4-Structures in Untranslated Regions. *Nat. Commun.* **2019**, *10* (1), 2421. https://doi.org/10.1038/s41467-019-10432-5.

(61) Bosco, B.; Rossi, A.; Rizzotto, D.; Hamadou, M. H.; Bisio, A.; Giorgetta, S.; Perzolli, A.; Bonollo, F.; Gaucherot, A.; Catez, F.; Diaz, J.-J.; Dassi, E.; Inga, A. DHX30 Coordinates Cytoplasmic Translation and Mitochondrial Function Contributing to Cancer Cell Survival. *Cancers (Basel).* **2021**, *13* (17), 4412. https://doi.org/10.3390/cancers13174412.

(62) Zhang, X.; Wigley, D. B. The “glutamate Switch” Provides a Link between ATPase Activity and Ligand Binding in AAA+ Proteins. *Nat. Struct. Mol. Biol.* **2008**, *15* (11), 1223–1227. https://doi.org/10.1038/nsmb.1501.

# CHAPTER 5. CONCLUSIONS

The RgpB protease is a molecular target of high pharmaceutical interest in the race for treatments against Alzheimer's disease. A detailed knowledge of its mechanism of action and inhibition is required to exploit its potential in drug design. The central objective of this thesis was to explore and apply different computational strategies to unravel the atomistic details of the catalytic and inhibition mechanisms of the RgpB protease. Overall, we were successful in i) elucidating the reaction mechanism by which RgpB catalyzes protein hydrolysis, ii) characterizing the inhibition process of RgpB in both covalent and non-covalent terms, and iii) successfully applying and testing diverse statistical thermodynamics methods based on the use of QM/MM potentials for the study of biological systems, i.e. umbrella sampling methods and metadynamics.

In this manner, and according with the proposed objectives, the computational studies carried out during the course of this thesis leaded to the following conclusions:

- Initially, in order to perform the study of the reaction mechanism catalyzed by RgpB, different protonation states of the catalytic dyad (Cys/His) were explored by performing classical MD simulations. Analysis of these simulations revealed conformational changes of a flexible loop that determines the stability of the interactions between the peptide and residues Glu152 and His211. An open conformation of the flexible loop leads to a loss of the reactive arrangement of the active site. On the other hand, a closed conformation that maintains the interactions between residues Glu152 and His211 and the peptide was identified as the reactive one, with protonated Cys244 and deprotonated His211.

  The most likely mechanism proceeds in two steps for the acylation step and one step for the deacylation step. The acylation and the deacylation steps have similar activation barriers, 23.4 kcal·mol$^{-1}$ and 23.5 kcal·mol$^{-1}$, respectively. Thes values were very close to the value deduced from previous experimental kinetic reports, 22.8 kcal·mol$^{-1}$, validating our results.

- Notably, the unusual configuration of the catalytic dyad (Cys/His) in the active site positions the peptide so that it interposes itself between the catalytic residues. Given this configuration, the role of the His211 residue does not coincide with other cysteine proteases deprotonating the catalytic cysteine. In this case, it plays an essential role by orienting the water molecule that hydrolyzes the acylenzyme via a water-mediated hydrogen bond. Furthermore, our findings highlight the crucial role played by the substrate in the activation of Cys244. This role should be taken into account for the future design of RgpB inhibitors for the treatment of Alzheimer's disease and other human conditions.

- In order to study the RgpB inhibition, six irreversible inhibitors patented as candidates for the treatment of Alzheimer's disease were simulated in complex with RgpB by classical molecular dynamics in their non-covalently bound state. Over these trajectories, binding enthalpies, interaction entropies, and binding free energies were estimated using MM/PBSA and alchemical transformations.

  The relative contact map between inhibitors and neighboring residues revealed that three of them (IH1, IH3 and IH5) present a similar pattern of interactions, while the other three (IH2, IH4 and IH6) which present lower binding energies, did not present consistent interactions with the key residues Ser213 and Glu214. The aforementioned interactions, showed favorable interaction energy when present. Thus, these analyses of the non-covalent complexes suggest that IH1 and IH3 would be the most promising candidates for future refinements.

- The reaction mechanism of the covalent bond formation between the inhibitors and the enzyme was computed by QM/MM MD simulations on IH3, which showed the highest binding free energy in the non-covalent complex. The resulting FES describes the reaction through a single concerted step, mediated by a water molecule. The reaction proceeds through an activation barrier (14.2 kcal·mol$^{-1}$) significantly lower to that previously reported for the proteolysis reaction (23.5 kcal·mol$^{-1}$). Thus, it is remarkable the ability of RgpB to react without the presence of highly reactive groups, enabling to exploit the re-design of covalent inhibitors, recognized for their potency, without predicting problems of selectivity. These results qualify RgpB as a pharmacological target that promises effective treatments by the use of these kind of inhibitors without side effects. Our study suggests that more potent inhibitors should contain hydrogen bond donor and acceptor groups to be able to interact with Glu214, Ser213, Asp158 and Asp281, and conserving the reported guanidinium group and warhead to provide potency and selectivity. The interactions reported here and the mechanistic details represent a key starting point for future re-design of prospective and efficient inhibitors for the treatment of Alzheimer's disease.

- Finally, in order to test other QM/MM statistical thermodynamics methods to study the complete landscape of the reactivity and inhibiton of RgpB, we explore the use of metadynamics. The mechanism of RNA-driven ATP hydrolysis of Prp2 was used as a benchmark in this study. The results showed that the ATP hydrolysis mechanism proceeds in four steps with the rate determining one being the attack of the nucleophilic water to the P$\gamma$:ATP and a free energy barrier in agreement with experimental kinetic data. It was also revealed a marked increase in the activation free energy barrier of the rate determining step in the absence of the RNA strand, confirming the dependency of the catalytic power in RNA.

Complementary, MD simulations disclosed the molecular terms of RNA-driven activation of ATP hydrolysis. An additional bioinformatic analysis reveals that the RNA-driven ATPase activation is conserved only across the DExH-box RNA helicases which, besides participating in splicesome remodeling, take part in other key cellular processes. It was suggested that in DExH-box helicases, an expanded RR-EE motif plastically regulates the ATPase/helicase function in response to the binding of an effector RNA strand. Thus, our findings provide a set of experimentally testable hypotheses to be carried forward in future studies aimed at deciphering the mechanism of DExH-box helicases.

These results show the versatility and efficiency of metadynamics in enzyme reactivity studies, suggesting that they can be successfully applied in a future research to the RgpB enzyme.

# APPENDIX A. LIST OF ABREVIATIONS

AM1 - Austin Model 1

ADP - Adenosine Diphosphate

ATP - Adenosine Triphosphate

B88 – Becke/1988 Functional

BFGS – Broyden-Fletcher-Goldfarb-Shanno

BLAST - Basic Local Alignment Search Tool

CNDO - Complete Neglect of Differential Overlap

CV – Colective Variable

DFT – Density Functional Theory

DNA - Deoxyribonucleic Acid

E – Enzyme (Not to be confused with the energy variable E)

ES – Enzyme:Substrate Complex

ESI – Enzyme:Substrate:Inhibitor Complex

FES – Free Energy Surface

FF – Force Field

FFT - Fast Fourier Transforms

GBSA - Generalized Born and Surface Area

GGA - Generalized Gradient Approximations

GPW – Gaussian and Plane Waves

GTH - Goedecker, Teter, and Hutter Pseudopotentials

HK – Hohenberg-Kohn

$IC_{50}$ - Half Maximal Inhibitory Concentration

IHC - Immunohistochemical

IUPAC - International Union of Pure and Applied Chemistry

IRC – Internal Reaction Coordinate

Kgp – Lysine Gingipain

LDA – Local Density Approximation

LJ - Lennard-Jones

LSDA - Local Spin Density Approximation

LYP – Lee–Yang–Parr Functional

MD – Molecular Dynamics

MM – Molecular Mechanics

MNDO - Modified Neglect of Diatomic Overlap

MO - Molecular Orbital

MTD - Metadynamics

NDDO - Neglect of Diatomic Differential Overlap

NPT- Isothermal Isobaric Closed Ensemble

NVT - Canonical Ensemble

P - Products

PBE – Perdew-Burke-Ernzerhof Functional

PBSA - Poisson–Boltzmann Surface Area

PCA – Principal Component Analysis

PDB – Protein Data Bank

PES – Potential Energy Surface

PM3 - Parametric Method 3

PM6 - Parametric Method 3

PME - Particle Mesh Ewald

QM/MM – Quantum Mechanics/Molecular Mechanics

S - Substrate

SASA - Solvent Accessible Surface Area

TS – Transition State (Not to be confused with the entropic term TS)

TST – Transition State Theory

Rgp – Arginine Gingipain

RMSD – Root Mean Square Deviation

RMSF- Root Mean Square Fluctuation

RNA - Ribonucleic Acid

WHAM – Weighted Histogram Analysis Method

ZDO - Zero Differential Overlap

# ANNEX A. ARTICLE 1

JCIM JOURNAL OF CHEMICAL INFORMATION AND MODELING

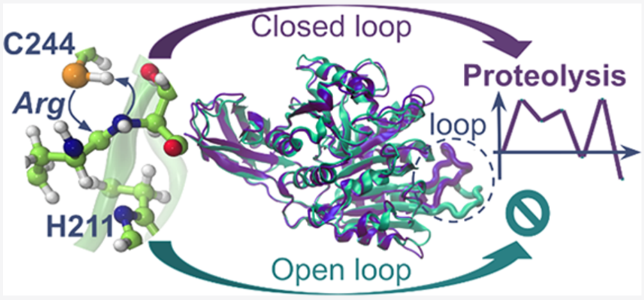
# Unrevealing the Proteolytic Activity of RgpB Gingipain from Computational Simulations

Santiago Movilla, Sergio Martí, Maite Roca,* and Vicent Moliner*

**Cite This:** *J. Chem. Inf. Model.* 2021, 61, 4582−4593

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information



**ABSTRACT:** Alzheimer's disease represents one of the greatest medical concerns for today's population and health services. Its multifactorial inherent nature represents a challenge for its treatment and requires the development of a broad spectrum of drugs. Recently, the cysteine protease gingipain RgpB has been related to neurodegenerative diseases, including Alzheimer's disease, and its inhibition appears to be a promising neuroprotective strategy. Given these features, a computational study that integrates molecular dynamics (MD) simulations with classical and hybrid quantum mechanics/molecular mechanics (QM/MM) potentials was carried out to unravel the atomistic details of RgpB activity. First, a preliminary study based on principal component analysis (PCA), determined the protonation state of the Cys/His catalytic dyad, as well as the crucial role of a flexible loop that favors reactive interactions of the catalytic residues and the peptide in the precatalytic state in its closed conformation. Then, different mechanisms were explored by means of QM/MM MD simulations. The most favorable mechanism consists of two stages. First is an acylation stage that takes place in two steps where, initially, the sulfur atom of the C244 residue attacks the carbonylic carbon of the peptide and the proton of the C244 residue is transferred to the amino group of the peptide in a concerted manner. Subsequently, the peptide bond is broken, and a fragment of the peptide is released. After that, the deacylation stage takes place in a single step where a water molecule attacks the carbonylic carbon of the peptide and a proton of the water is transferred to the C244 residue. The free energy barrier of the rate limiting step is in very good agreement with available experimental data. The mechanism exhibits an unusual role of H211 residue compared with other cysteine proteases but a crucial role of the peptide in triggering the catalysis. Notably, the atomic and energetic particularities found represent a significant contribution to the comprehension of the reaction mechanism and a great opportunity for the design of efficient inhibitors of gingipain RgpB.

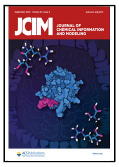
## 1. INTRODUCTION

Alzheimer's disease is one of the most challenging conditions for health research worldwide.[1,2] It is a progressive neurodegenerative brain disorder that causes devastating functional, cognitive, and behavioral problems.[1−4] Categorized as a multifactorial disease, treatment has been attempted without success across a broad spectrum of metabolic targets.[5] Thus, in the absence of a successful treatment, comprehension of the reaction mechanism of enzymatic systems related with Alzheimer's disease and a proposal of new targets as potential therapeutic strategies is necessary to bring this overwhelming disease under control.[1−5]

For some years now, several reports have been linking oral health problems to cognitive impairment conditions such as Alzheimer's disease. Specifically, infections with *Porphyromonas gingivalis*, the cornerstone pathogen in the development of periodontitis, have been shown to have a significant influence on the formation of amyloid $\beta$-peptide (A$\beta$) plaques, development of dementia, and aggravation of Alzheimer's disease.[6−11] Skillfully, a recent study identified in the brain of Alzheimer's patients a group of enzymes secreted by *Porphyromonas gingivalis*, gingipain proteases, as a direct cause of this neuronal damaging effect. The study concluded that gingipain small inhibitors have promising neuroprotective brain effects.[11]

Gingipains are enzymes belonging to the CD clan of cysteine peptidases. The CD clan involves a number of cysteine proteases within a wide range of parasitic protozoa. Members of this clan mainly differ from those of all other clans in terms of primary and tertiary structure, the proteins that they hydrolyze, and so their metabolic functions.[12] More in depth, two large groups of gingipains can be distinguished, Lys-gingipains (Kgp) and Arg-gingipains (Rgp), according to the residue they recognize at the P1 position to cleave peptides.[13]

**Received:** June 11, 2021
**Published:** September 2, 2021



https://doi.org/10.1021/acs.jcim.1c00666

Although all gingipains are related to a wide range of diseases from gingivitis to cardiovascular problems,[14−17] RgpB shows a stronger correlation with the progression of Alzheimer's disease.[11] RgpB is the gingipain-R encoded by the *rgpb* gene and differs from RgpA and HRgpA forms by the absence of the hemagglutinin/adhesin domains.[18−20] With this in mind, a better understanding of the action of this attractive pharmacological target is necessary in order to design possible efficient inhibitors to treat neurodegenerative diseases.

RgpB activity is extended to the hydrolysis of basically any peptide bond with arginine at the P1 position.[21] Some other residue preferences along the peptide chain have been reported, however, none as essential as that of arginine.[13,22] Despite the limited kinetic information available, it is well-known that RgpB exhibits proteolytic activity in a broad pH range between 6.5 (50%) and 9.5 (100%). Outside this range, it is hypothesized to become structurally unstable.[23] As a result of the aforementioned situation, RgpB is capable of destroying human connective tissues, cell surface proteins and receptors, cytokines, components of the coagulation and complement cascades, heme and iron-binding proteins, immunoglobulins, and proteinase inhibitors.[24−27] This fact enhances the pharmacological importance of understanding in depth the reaction mechanism of RgpB for exploiting medical proposes.[28]

Some features are common and accepted in the mechanisms that have been reported so far for almost all cysteine proteases reactions.[29−42] In general, the reaction proceeds through two main stages. The first stage, the acylation, corresponds to the formation of the acylenzyme through the nucleophilic attack of the cysteine sulfur and the displacement of the peptidic leaving fragment. Afterward, in a second stage, a base-activated water molecule attacks the C1 of the peptide, and the cysteine−peptide bond is broken to give way to the second product release. Despite these common features and other similarities between cysteine proteases, numerous studies have shown a remarkably wide reactivity among them. Disparity between cysteine proteases ranges from significant structural differences in active sites to small details in the transition states (TSs) of the rate-determining steps. In fact, these differences are the cornerstone in the design of potent and selective inhibitors/drugs. Our group's experience[38,40,43−45] and other studies on cysteine proteases[29−42] can be used to focus the attention on those critical points. Some of the most differentiated factors among the mechanisms of cysteine proteases include the protonation state of the Cys/His catalytic dyad, its role into the reaction mechanism, the residues that act as acids/bases, and whether the bond forming and bond breaking processes are concerted or stepwise.[29−42]

In the reaction mechanism of cysteine proteases, it is generally accepted that a proton transfer takes place between the Cys/His catalytic dyad prior to the formation of the acylenzyme.[29−42] However, in the particular case of RgpB, the peptide is located between these two catalytic residues, and direct proton transfer between the catalytic Cys and His residues does not look feasible.[22] Given this, few known mechanisms in proteases are viable to adapt to RgpB. Furthermore, E-64,[46] a commonly used reference inhibitor of cysteine proteases, shows no inhibitory activity on this cysteine peptidase family, as it does on all others.[47] This fact, together with the atypical spatial disposition of the catalytic dyad residues in RgpB and the peptide, point to crucial differences in the reaction mechanism that must be explored for the drug design process.

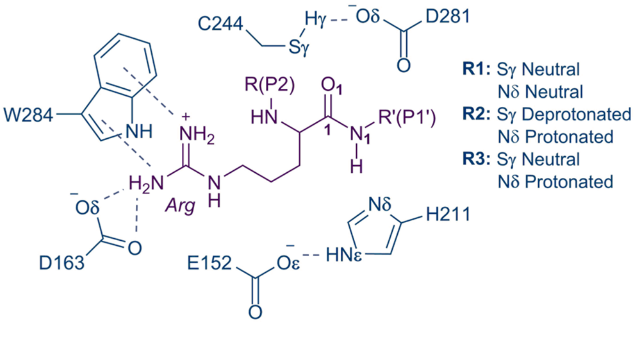
Herein, we conducted a computational study that integrates molecular dynamics (MD) simulations with classical and hybrid quantum mechanics/molecular mechanics (QM/MM) potentials and structural data analysis aimed to understand the activity of the RgpB gingipain protease from *Porphyromonas gingivalis* in atomistic detail. Our results show conformational changes of the precatalytic state depending on the protonation states of the catalytic dyad Cys/His. Additionally, from the most stable and reactive structure, we propose a molecular mechanism by which the proteolysis reaction can proceed, revealing new valuable particularities important for future rational drug design.

## 2. COMPUTATIONAL DETAILS

**System Set Up.** The initial coordinates of the system were obtained from the crystal structure of RgpB from *Porphyromonas gingivalis* in complex with a peptide-like inhibitor (PDB code 1CVR).[22] The inhibitor was replaced by the protein fragment Cys-Ala-Tyr-Arg-Thr-Ser-Pro (acetylated terminals) of human pancreatic ribonuclease (UniProtKB-07998)[48] preserving as many atoms of the crystallized inhibitor as possible. The fragment was chosen based on the inhibitor's size, its electrostatic neutrality, and reports of being hydrolyzed by RgpB.[23] The coordinates of the missing atoms and hydrogens were added in sterically favorable positions. Protonation states for the titratable residues were selected based on the results provided by the PROPKA3 software.[49] The selected pH value was 7.5 as adjusted experimentally in the activity and inhibition studies.[11,21,22] A total of 16 to 18 $Na^+$ ions,[50] depending on the protonation states of the Cys/His catalytic dyad, were added in electrostatically optimal positions (those positions where the electrostatic potential reaches a maximum) around the enzyme in order to neutralize the system. Finally, the system was solvated with a 107.5 Å$^3$ cubic box of water molecules (TIP3P)[51] with a minimum distance of 15 Å between any protein atom and the edge of the box. The water molecules from the crystal structure were preserved. The complete system contains ∼116 k atoms. All of the steps previously described were done using the AmberTools17 tleap package.[52]

As mentioned in the Introduction, the peptide is located between the catalytic dyad, and no proton transfer is possible between C244 and H211 residues (see Scheme 1 and Figure 1). Moreover, in the absence of experimental data, three initial structures with different protonation states of these

Scheme 1. General Disposition of the Active Site and Protonation States of the Proposed Initial Structures[a]



[a] Interactions between the residues and the peptide are displayed in dashed lines.

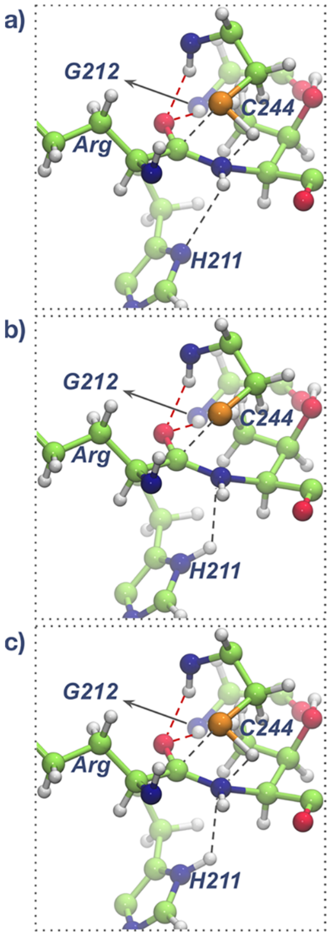
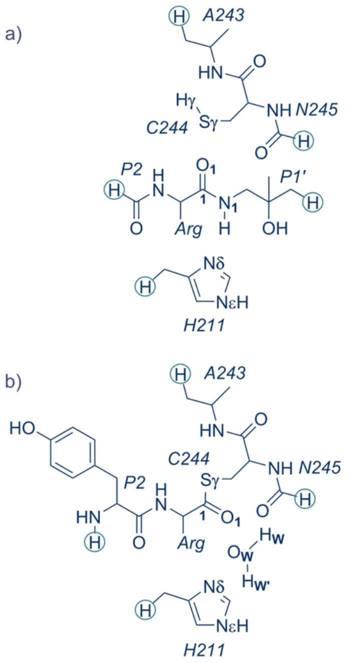


**Figure 1.** Initial structures of the considered precatalytic states (a) **R1**, (b) **R2**, and (c) **R3**. Adapted and prepared from the crystal structure template, 1CVR.[22] The interactions between the peptide and the backbone of oxyanionic hole residues, C244 and G212, are shown in red dashed lines, while the possible bond forming and bond breaking distances are in black dashed lines.

two residues were generated and classically simulated to evaluate the stability and reactivity of the precatalytic state because of the possible high impact over the reaction mechanism. Scheme 1 shows the structures of all plausible and chemically reasonable combinations of protonation states of C244 and H211 residues (**R1**, **R2**, and **R3**). The first one has both residues neutral (**R1**). The second one corresponds to its protomer with the C244 residue deprotonated (negatively charged) and the H211 residue protonated (**R2**). The last one shows the C244 residue neutral and the H211 residue protonated (**R3**). There is another protonation state, the C244 residue deprotonated and H211 neutral, but it does not have chemical sense for the mechanistic point of view because none of the residues could act as an acid. Thus, this last protonation state was discarded. Frequently, in cysteine proteases the first two protomers (**R1** and **R2**) are in chemical equilibrium[35,38,40−42] and the preference for one over the other can be explored by QM/MM methods. However, as previously mentioned, RgpB has the peptide between these residues, and a proton exchange between C244 and H211 is physically inconceivable. The last protonation state (**R3**) is uncommon but was previously proposed[22] for RgpB, so we decided to evaluate it. Scheme 1 also displays some key residues that interact with the peptide, C244 or H211 residues. Figure 1 shows the initial structures of the three different precatalytic states to be studied (**R1**, **R2**, and **R3**) with the interactions with the oxyanionic hole residues, G212 and C244, and with H211.

**Equilibration of Initial Reactants States.** Optimizations and classical MD simulations were carried out to relax and equilibrate **R1**, **R2**, and **R3** initial states in solution using AMBERff14SB[53] and TIP3P[51] force fields to describe the protein and solvent water molecules, respectively. First, a minimization was performed on the solvent molecules, ions, and hydrogens using 2500 minimization steps with the conjugate gradient algorithm. Then, a short 2 ns dynamic of the solvent molecules and ions with the constrained position of the heavy atoms of the protein (restraint constant of 300 kcal·$\text{mol}^{-1}\cdot\text{Å}^{-2}$) was carried out, followed by two energy minimizations, one with the protein backbone restrained and another completely unrestrained. Then, the whole system was heated in four consecutive dynamics. The heating dynamics were as follows: (1) NPT ensemble, 100 K and 1 bar, time step of 0.5 fs; (2) NVT ensemble, 200 K, time step of 0.5 fs; (3) NVT ensemble, 300 K, time step of 0.5 fs; and (4) NVT ensemble, 300 K, time step of 2 fs. Subsequently, 50 ns of NVT was run at 310 K. Up to this point, restraints were applied to interactions that we considered crucial for the catalysis (E152−H211, C244−peptide, H211−peptide, D281−C244). Finally, sampling production dynamics were performed with the system completely unrestrained with the NVT ensemble at 310 K and a time step of 2 fs. In particular, 150 ns of classical MD were followed by 2 ns of QM/MM MD using the semiempirical PM6[54] Hamiltonian for the quantum region and the AMBERff14SB[53] and TIP3P[51] force fields to describe the protein and the water molecules, respectively. The QM/MM frontier was treated using the link atom procedure. Atoms included in the quantum region are shown in Scheme 2. The

**Scheme 2. Schematic Representation of the Region Treated Quantum Mechanically to Explore the Mechanism of (a) Acylation Stage and (b) Deacylation Stage (Link Atoms Added to Treat QM−MM Frontiers Depicted in Circles)**



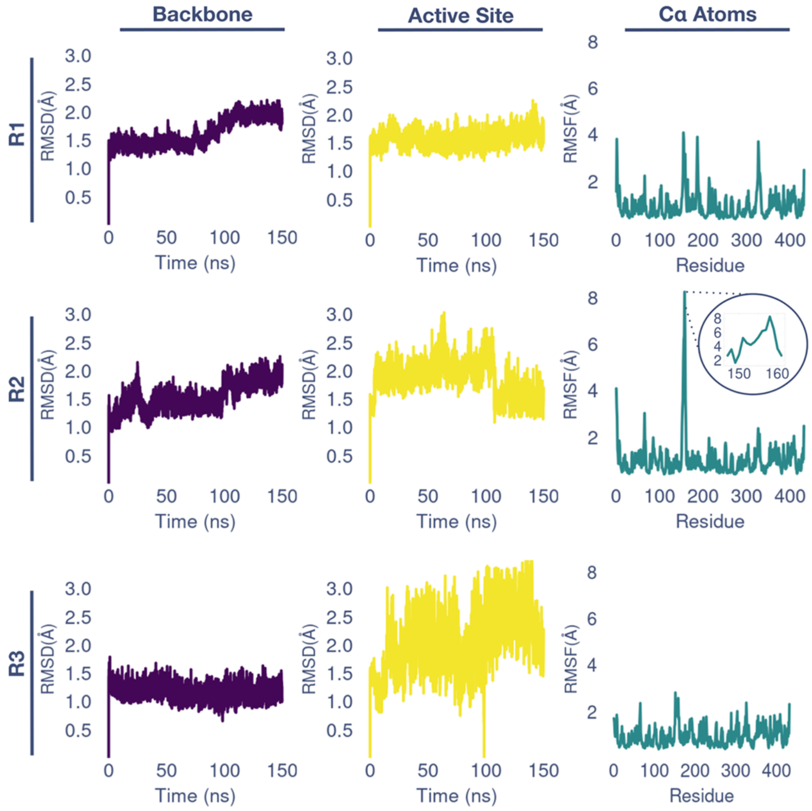


**Figure 2.** Geometrical analysis of the trajectories obtained for **R1**, **R2**, and **R3** systems along the 150 ns of classical MD simulations. Left panels: time evolution of the root-mean-square deviation (RMSD) of protein backbone atoms (Cα, C, N, O). Center panels: time evolution of the RMSD of the active site atoms (heavy atoms of the E152, H211, C244, D281 residues and peptide). Right panels: root-mean-square fluctuation (RMSF) analysis of protein α carbon atoms. The inset in the RMSF plot of the **R2** system shows the values of RMSF of the flexible loop 147−159 residues.

classical calculations were run in the AMBER GPU[55,56] software version. The cutoff limits for short-range nonbonded interactions were 10 Å, and a Particle Mesh Ewald (PME)[57,58] model was used for the long-range interactions. Temperature control was performed using Langevin dynamics[59,60] with a 3 $ps^{-1}$ collision frequency. For all equilibration simulations, the SHAKE[61,62] algorithm was used to constrain light atoms, and the velocity Verlet[63] algorithm was used to update the velocities. Equilibrium convergence was confirmed by the evaluation of the RMSD of the backbone atoms using the CPPTRAJ package.[64]

**Potential of Mean Force (PMF).** From QM/MM equilibrated structures, 1D and 2D potential energy surfaces (PES) were calculated through sequential minimizations along selected collective variables that best describe each chemical transformation. A conjugate gradient algorithm was employed for the minimizations using a gradient tolerance of 0.1 kcal $mol^{-1}$ as a convergence criterion. Later, free energy surfaces (FESs) were generated in terms of PMFs[65,66] at 310 K using the structures of the above-mentioned 1D and 2D PESs as starting points for each window. Each window had a relaxation time of 5 ps and a sampling time of 25 ps using a time step of 0.5 fs in NVT ensemble. Temperature control was performed using Langevin dynamics[59,60] with a 3 $ps^{-1}$ collision frequency. The umbrella sampling[67] method was used to restrain the reaction coordinates. The force constant used for each window was 580 $kcal\cdot mol^{-1}\cdot Å^{-2}$, and the window width was between 0.05 and 0.1 Å depending on the distinguished reaction coordinate. The number and the width of the windows selected ensure a correct overlapping of windows. The umbrella integration method,[68] as implemented in the QM3 suite,[69] was used to analyze the biased sampling dynamics and to generate the PMFs along selected coordinates. The reaction coordinates chosen for each step, the number of windows, and the intervals of the reaction coordinates are described in Table S1 of the Supporting Information. In order to improve the PM6 energy function used to generate the PMFs, an electronic correction has been applied over the PM6-level optimized stationary points. Transition state structures were selected by density peaks clustering processes[70] (RMSD cut 0.7 Å) at the saddle points of the PM6/MM FESs. Subsequently, these structures were optimized and verified by analysis of the Hessian and by tracing down the intrinsic reaction coordinate path (IRC), and the extremes were finally optimized. Single point energy calculations were carried out on these structures at the DFT/MM level, employing the PBE functional[71−74] and the D3(BJ) dispersion correction[75] with the 6-311+G** basis set, to correct the PM6/MM electronic energy differences. The thermal contributions calculated by the statistical methods at the PM6/MM level were thus preserved. In order to verify the results, the transition states corresponding to the rate limiting steps of both reaction stages (acylation and deacylation) were also optimized using the high level method and verified by inspection of the normal modes. Charges were calculated from electrostatic potentials using a grid-based method (CHelpG)[76] on the stationary structures localized using the PBE+D3(BJ)/MM method with the 6-311+G** basis set. All of these calculations were carried out using Amber14,[77] Gaussian 09,[78] and fDynamo[79,80] software integrated with the QM3 suite.[69]

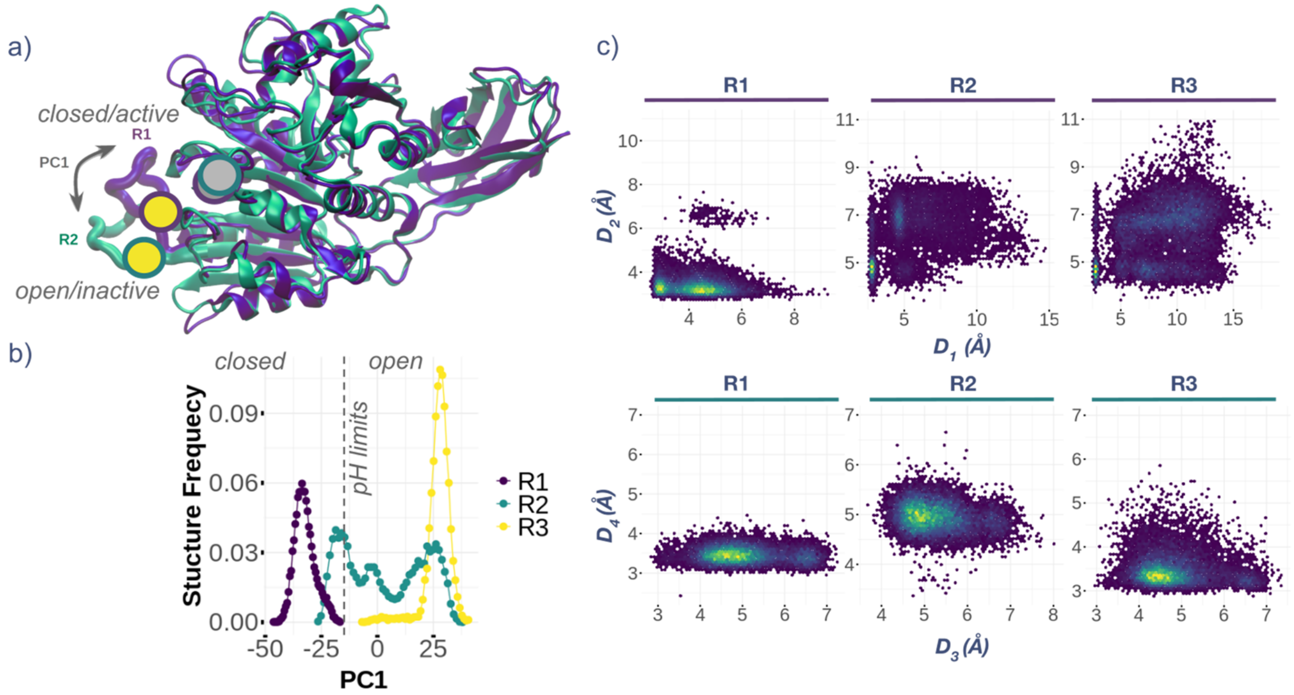


**Figure 3.** (a) Overlay of the protein structures with the open (**R2**) and closed (**R1**) conformations of the 147−159 loop defined by the PC1. Gray and yellow circles represent H211 and E152 residues, respectively. (b) Structure frequency of each reactive structure over the PC1 for the open and closed conformations of the 147−159 loop. (c) 2D representations of some distances between the Cys/His catalytic dyad, the peptide, and other residues. D1, [Nε:H211−Oε:E152]; D2, [Nδ:H211−N1:Pep]; D3, [Sγ:C244−Oδ:D281]; D4, [Sγ:C244−O1:Pep]. For D1 and D3, the distances were considered to the closest oxygen atom of the carboxylic group. Distances are in Å.

## 3. RESULTS AND DISCUSSION

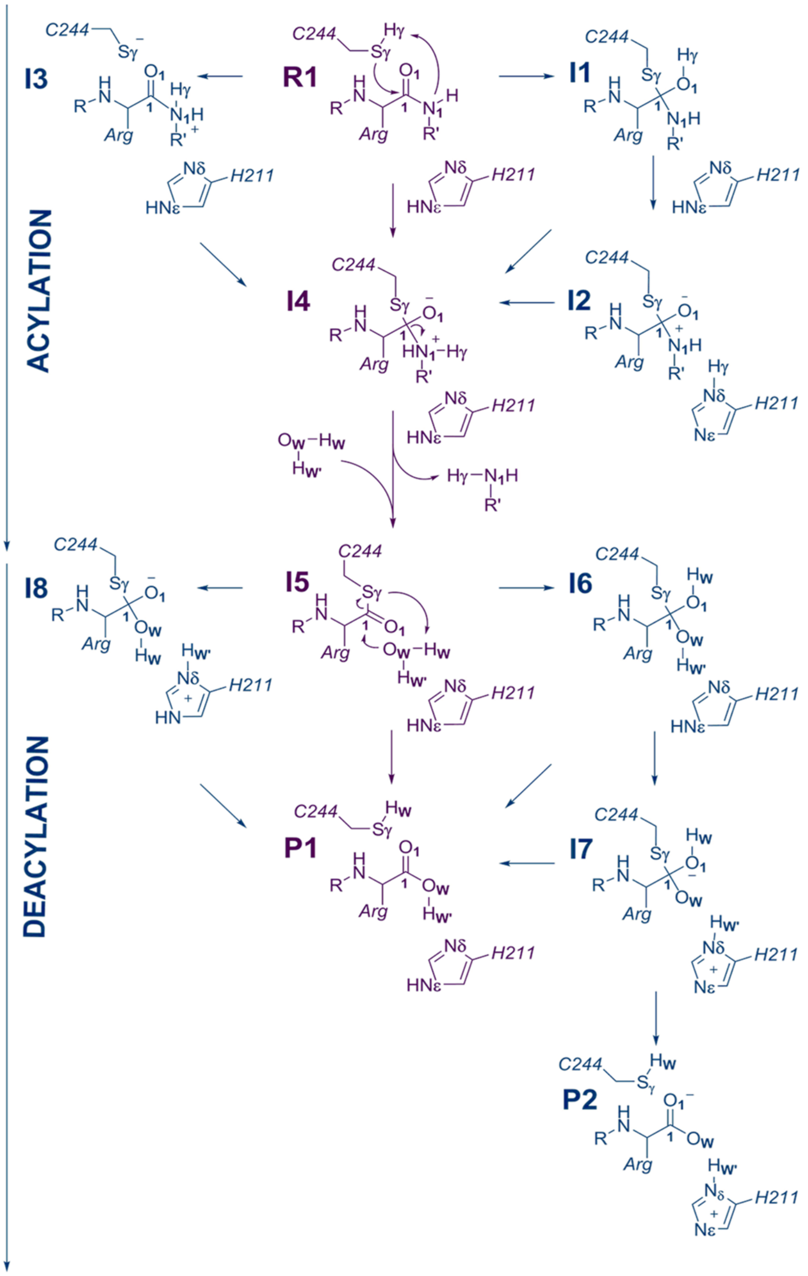
**Precatalytic State—Protonation States Strongly Affect Structural Stability and Catalytic.** *Power of RgpB.* The first step in our study consists of determining the structure and the protonation state of the catalytic dyad H211 and C244 in the initial enzyme−substrate complex based on classical MD simulations. The analysis of the time evolution of the root-mean-square deviation (RMSD) graphs of the protein backbone obtained with the three possible systems, **R1**, **R2**, and **R3** (see Scheme 1), shows how the three systems reached equilibrium, and none of the systems exceeded, on average, the 2.0 Å threshold of the crystal resolution (see Figure 2). Nevertheless, remarkable differences can be detected. Thus, while the RMSD computed for all of the backbone atoms of the protein are similar in the three systems, the RMSDs computed on the atoms of the active site for systems **R2** and **R3** show larger relative standard deviations than for system **R1** (see Table S2 in the SI). Moreover, the root-mean-square fluctuation (RMSF) analysis of the protein $\alpha$-carbon atoms in **R2** shows higher fluctuations in the loop between 147 and 159 residues (see the inset graphic of the RMSF of **R2** system). The significantly high fluctuations observed in the **R2** system and larger deviations in the RMSD in the active sites of **R2** and **R3** systems, suggest that the system with both neutral residues (**R1**) must be the most stable.

In order to get a deeper insight into these fluctuations, a principal component analysis (PCA) was carried out, revealing two possible conformations, open and closed, of loop 147−159, containing the catalytic residue E152 that interacts with the H211 residue. The first component PC1 reflects the opening and closing movement of this loop as shown in Figure 3a. Figure 3b shows how, while **R3** solely explores open conformations, **R2** varies across a much wider range of values on PC1 and **R1** was restricted to closed conformations during the whole simulation. Conformational changes of the loop between open and closed conformations in **R2** is correlated with the highly fluctuating regions in the RMSF (right panel Figure 2) while the large frequencies of **R1** in the closed conformation and **R3** in open conformations (Figure 3b) are in agreement with few changes in the loop conformations (low RMSF values).

A detailed structural analysis showed that the position of this loop can affect the interactions between E152, H211, and the peptide (Figure 3a). So, in the open conformations, these catalytic residues are not constantly close to the active site, and thus the proteolytic activity must be decreased. H211 has been reported to be essential for catalysis, but the only role proposed so far is as an acid at the beginning of the catalysis.[22] This proposed role would force the histidine to be protonated at the beginning of the reaction. However, simulations with protonated histidine showed significantly large fluctuations over some critical distances and an open conformation of the loop (**R3**) or a wide variety of conformations (**R2**), giving rise to inactive structures.

As can be seen by the population analysis of the structures generated along the MD simulations (see Figure 3c), the Nδ:H211−N1:Pep distance was kept above 6.5 Å for most of the time during the **R2** and **R3** simulations. In contrast, when the H211 residue was neutral (**R1**), the distance between these two nitrogen remained stable at an average of 3.2 Å during the MD simulations of the **R1** system. This suggests that the structural repercussion of opening this loop is stronger than the hydrogen bonds or electrostatic interactions that may exist at the active site when H211 residue is charged. As expected, also, the distance between Oε:E152 and Nδ:H211 is strongly influenced by the opening of the loop. The **R2** and **R3** systems

**Scheme 3. Explored Molecular Mechanisms for the Proteolysis Reaction Catalyzed by the RgpB Gingipain**[a]



[a]The most plausible mechanism, according to the QM/MM free energy profiles, is depicted in purple.

show very poor interactions between these two residues, while **R1** remains stable in a reactive arrangement. The interaction between the peptide with C244 residue is also affected by the protonation state of C244 and H211 residues. Particularly, while the **R1** and **R3** systems have an average distance of 3.4 Å between S$\gamma$:C244 and C1:Pep atoms, the **R2** system places the sulfur at an average distance of 5.0 Å due to a rotation of the thiolate group pointing to the opposite side of the peptide. Our predicted reduced catalytic power of the enzyme due to conformational changes induced by the presence of the thiolate ion (**R2** structure) is in agreement with a previous hypothesis based on experimental results.[47] An analysis of the hydrogen bond population involving the proton of the thiol group of C244 residue showed a strong interaction with the O1 atom of the electrophilic carbonyl group of the peptide. This interaction locates the sulfur close to the carbon. This also means that the interaction between D281 and C244 residues, that was assumed to be important for the mechanism,[22] is not observed in any of the simulations. In fact, the D281 residue was far from the C244 residue in all of the simulations we performed. Additionally, the absence of a basic residue (like D281) close to the cysteine residue in the active site of other crystal structures[21] of gingipains allows for a conclusion that this residue is not essential for the proteolysis reaction. Previous computational studies[81] also support this structure by showing that in solvent accessible locations the interaction Cys:S−H···$^{-}$O−C:Asp is less stable than Wat:O−H···$^{-}$O−C:Asp.

Finally, the calculated p$K_a$ values for C244 and H211 residues in the open conformation are 6.31 and 9.92, respectively. These values closely match the p$K_a$ experimentally[47] estimated not only for RgpB but also for other related gingipains. Still fascinating is the fact that these values coincide with the pH limits from which catalytic activity decreases dramatically, doubtless because of structural instability. In light of our findings, such structural instability may be related with the open/closed conformations of the loop associated with the Cys/His protonation states. At pH values between 6.5 and 9.5, the enzyme has both residues neutral, and it is in closed (active) loop conformation.

In all, our simulations pointed to the H211 and the C244 residues in their neutral states (**R1**) as the most probable precatalytic state. The other two protomers (**R2** and **R3**) proved to be significantly more unstable and thus implausible as starting point structures. This result agrees with the protonation states proposed by Elsässer and co-workers for the cysteine protease legumin, which shows a similar spatial distribution of the catalytic Cys/His dyad and the peptide.[39]

As revealed by analysis of the average structure of R1, the peptide is in between the catalytic Cys/His dyad (see Figure S1). The thiol group of the C244 residue is oriented to the peptide. Moreover, the hydrogen bond interaction between E152 and H211 residues is kept; the flexible loop is in a closed conformation. With regard to the arginine residue of the peptide, a salt bridge interaction is shown with the D163 residue and a $\pi$ staking interaction with the W284 residue.

*Acylation Stage Mechanism—Essential Substrate Assistance Triggers the Proteolytic Reaction.* Starting from the **R1** system, that has been demonstrated to be the most probably reactant state, several mechanisms were proposed and studied as shown in Scheme 3. All of the PMFs are displayed in Figure S2 in the Supporting Information, and the free energy profile of the most feasible reaction mechanism is depicted in Figure 4. The stationary point structures of the most plausible mechanisms are depicted in Figure 5. The free energy profiles of the alternative reaction mechanisms are shown in Figure S3 in the Supporting Information. The first mechanism explored involves the proton transfer from the Sγ:C244 atom to the O1 atom of the peptide (O1:Pep) that appears to take place concomitant with the attack of the Sγ:C244 atom on the C1:Pep atom to achieve an **I1** intermediate. The associated transition state **TS$^{\mathbf{R1\text{-}I1}}$** has a barrier of 24.5 kcal·mol$^{-1}$ (see Figure S3 in the SI), and the intermediate **I1** has an energy of 15.6 kcal·mol$^{-1}$ over **R1**. Starting from **R1** or **I1**, there are alternative paths to reach **I4**. One of them, from **I1**, consists of the proton transfer from the O1:Pep atom to the Nδ:H211 atom (**I1** → **I2**), which would serve as a mediator to transfer the proton to the N1:Pep atom in a subsequent step (**I2** → **I4**). The free energy barrier to reach the **TS$^{\mathbf{I1-I2}}$** transition state is 29.3 kcal·mol$^{-1}$, while **I2** is found at 28.7 kcal·mol$^{-1}$ above **R1** (see Figure S3 in the SI). This high free energy barrier involving H211 as a mediator of proton transfer reflects a low basicity of the Nδ:H211 atom at this point in the reaction. To avoid crossing through this high energy, we studied the possibility of a direct transfer from the O1:Pep atom to the N1:Pep atom (**I1** → **I4**). However, as expected, formation of a four-membered transition state implies a substantial energy penalty, 46.5 kcal·mol$^{-1}$ over **R1** and 30.9 kcal·mol$^{-1}$ over **I1**. A similar free energy barrier, 32.7 kcal·mol$^{-1}$, was earlier reported for an analogous process in another cysteine protease.[38] Thus, from the **I1** intermediate onward, there is no viable mechanism for the reaction to proceed.

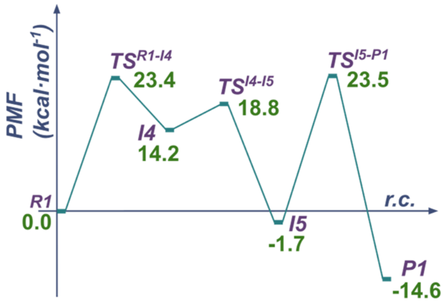


**Figure 4.** PBE+D3(BJ):PM6/MM free energy profile of the proteolysis catalyzed by the RgpB gingipain.

Then, another possible mechanism, starting from **R1**, consists of the proton transfer from the Sγ:C244 atom to the N1:Pep atom (**R1** → **I3**) and subsequently the attack of the Sγ:C244 atom on the C1:Pep atom (**I3** → **I4**). However, despite our efforts, the **I3** alternative intermediate with the charged Sγ:C244 atom and unbound from the C1:Pep atom could not be localized on any of the computed surfaces. Instead, the concerted reaction mechanism from the **R1** to **I4** intermediate appears to be viable, with a free energy barrier of 23.4 kcal·mol$^{-1}$ and the **I4** intermediate 14.2 kcal·mol$^{-1}$ over **R1** (see Figure 4). The nucleophilic attack of the Sγ:C244 atom on the C1:Pep atom is facilitated by the oxyanion hole formed by the backbone nitrogen protons of G212 and C244 residues. The localized transition state (**TS$^{\mathbf{R1\text{-}I4}}$**) shows an advanced state of proton transfer, which is confirmed by analysis of the free energy surface (Figure 5 and Figure S2 in the SI). This is also reflected in the increase of the negative charge located on the Sγ:C211 atom (see Table S3 in the SI) and the increase of the positive charge on the Hγ:C211 atom from reactants (**R1**) to the transition state (**TS$^{\mathbf{R1\text{-}I4}}$**). On the other hand, the charge on the electrophile C1:Pep atom increases while the charge on N1:Pep atom decreases as the reaction proceeds. The endergonic character of this step is determined by the pseudo-stability of the **I4** structure, in which the N1:Pep−C1:Pep and the Sγ:C211−C1:Pep bond distances are significantly longer than the standard values (1.61 and 2.04 Å, respectively). In fact, decomposition of **I4** into the **I5** intermediate by the N1:Pep−C1:Pep peptide bond breaking occurs through a low barrier of 4.6 kcal·mol$^{-1}$. Intermediate **I5** is located at −1.7 kcal·mol$^{-1}$ with respect to **R1**, a minimum that is both structurally and energetically similar to that reported by Elsässer and co-workers (Figures 4 and 5).[39] Thus, from all of the possible explored mechanisms for the acylation step (**R1** to **I5**), the most favorable reaction path takes place in two steps, through a zwitterionic stable intermediate, **I4**. The 23.4 kcal·mol$^{-1}$ barrier associated with the **TS$^{\mathbf{R1\text{-}I4}}$** transition state is in good agreement with the experimental data of the 22.8 kcal·mol$^{-1}$ reported value.[21] A picture of the rate limiting **TS$^{\mathbf{R1\text{-}I4}}$** optimized at the PBE+D3(BJ)/MM level is shown in Figure 6.

The role of the N1:Pep atom as the base has been previously proposed in a similar mechanism to that of the cysteine protease legumain.[39] Then, this reaction catalyzed by RgpB gingipain can be classified as a Substrate Assisted Catalysis (SAC).[82] This role of the substrate can also explain the inability of compound E-64[46] to inhibit this family of enzymes.[47] In E-64,[46] the group that should act as a basic/leaving group and the carbonyl group are distant as a consequence of the stereochemistry of the reactive epoxide.

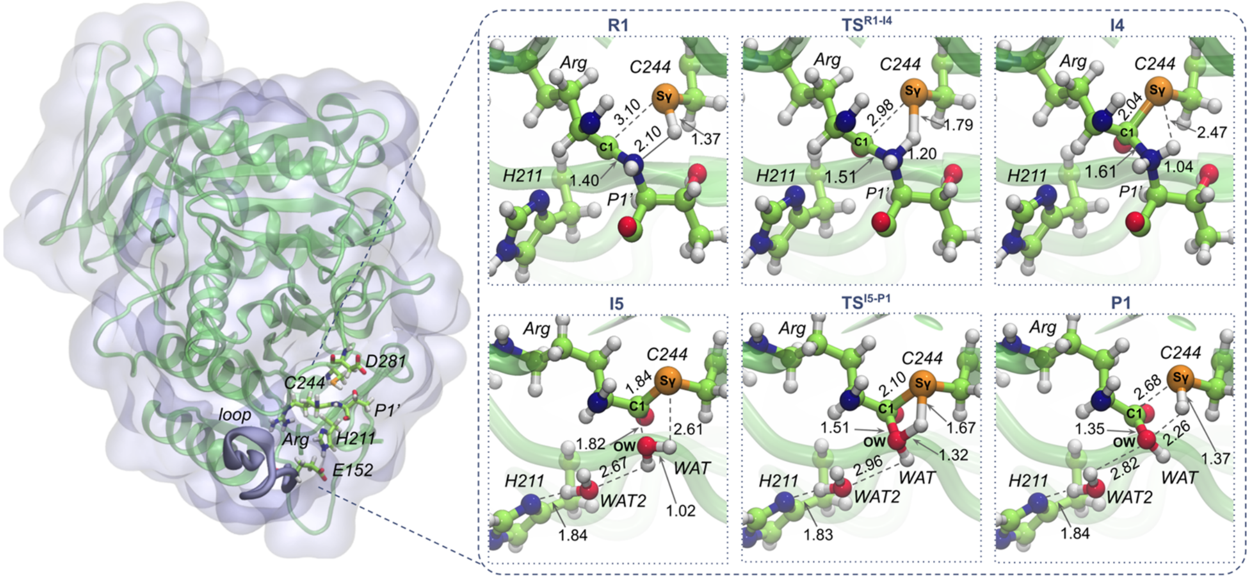


**Figure 5.** Left panel: Representation of the whole protein where key residues and the peptide are represented as sticks. Right panel: PM6/MM optimized structures of the key states appearing along the most favorable proteolysis reaction mechanism: **R1**, **TS**$^{\mathbf{R1\text{-}I4}}$, **I4**, **I5**, **TS**$^{\mathbf{I5-P1}}$, and **P1**. Distances are in Ångstroms.

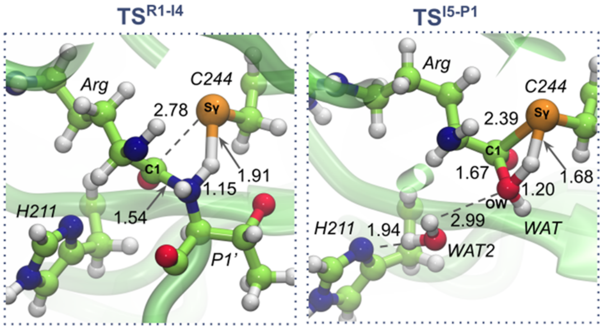


**Figure 6.** PBE+D3(BJ)/MM optimized structures of the rate limiting transition states of the acylation (**TS**$^{\mathbf{R1\text{-}I4}}$, left panel) and deacylation (**TS**$^{\mathbf{I5-P1}}$, right panel) stages in the reaction of proteolysis catalyzed by the RgpB gingipain.

Thus, the relative distribution is inappropriate to carry out the attack with the sulfur atom of C244 residue that is not enough acid to be previously deprotonated. Then, the presence of a correctly disposed basic/leaving group must be carefully considered when designing covalent inhibitors for gingipains. In this regard, due to the values of activation and reaction free energies of the **R1** → **I1** step, similar to **R1** → **I4** step, new inhibitors could be designed mimicking this mechanism.

We also proceed to analyze the interaction energies between the protein and the peptide in this first step (see Figure S4). The results show strong interactions between the peptide and D163 and W284 residues. This result was expected given that these residues are responsible for the arginine selectivity of the RgpB enzyme. D163 and W284 residues strongly interact with the guanidinium group of arginine via a salt bridge and $\pi$ stacking interactions (see Scheme 1 and Figure S1), respectively, making the enzyme highly selective for arginine-containing peptides. It is also noteworthy that H211 and E152 residues also interact strongly with the peptide (see Scheme 1 and Figure S1). These results reveal an exceptional role of these residues in the electrostatic peptide−enzyme interaction. Further, they are more determinant in the electrostatic stabilization than even the oxyanionic hole residues, C244 and G212, which exhibit lower interaction energies. That means that H211 and E152 residues play an essential role in the structural/electrostatic stability of the precatalytic complex (see Figure S1).

*Deacylation Stage—An Uncommon Role of Histidine Residue.* Once the acylation step is completed, two water molecules occupy the space generated by the release of the first fragment of the peptide. An analysis of the previous stationary point structures of the mechanism revealed that these two water molecules are able to enter once the first fragment of the peptide has been released. No water molecules were present around the C1:Pep atom prior to the **I5** intermediate. Several mechanisms were proposed for the deacylation stage (see Scheme 3). On the basis of a previous study,[39] we considered the concerted attack of a water molecule to the C1:Pep atom together with the proton transfer to the O1:Pep atom (**I5** → **I6**). Although this process leads to a stable intermediate (−5.1 kcal·mol$^{-1}$), it has a high activation energy, 32.9 kcal·mol$^{-1}$ (**TS**$^{\mathbf{I5-I6}}$, see Figure S3 in the SI). From **I6**, two different mechanisms can be explored to obtain products. On one hand, the reaction mechanism can take place through a direct proton transfer from the O1:Pep atom to the S$\gamma$:C244 atom to reach the **P1** protomer. The activation free energy (**TS**$^{\mathbf{I6-P1}}$) of this step is 14.2 kcal·mol$^{-1}$. On the other hand, **I6** can evolve into an alternative product with the protonated histidine (**I6** → **I7** → **P2**). The mechanism results in an almost barrierless process, where the highest free energy barrier corresponds to the step from **I7** to **P2** with an activation energy of 0.8 kcal·mol$^{-1}$. The supposed product obtained by this mechanism has a particular low energy, a fact that can be rationalized due the salt bridge formed between the resulting two charged residues (peptide and H211 residue). Anyway, none of these reaction mechanisms are feasible, considering the first high free energy barrier (**TS**$^{\mathbf{I5-I6}}$).

The role of the H211 residue as the base to activate the water molecule to attack the C1:Pep atom to reach the **I8** intermediate was also explored. Due to the conformation of the active site, once the H211 residue is protonated by a water molecule, it cannot transfer the proton to the C244 residue to reach **P1**. Instead, the proton from the hydroxylic group bonded to C1:Pep can be transferred to Sγ:C244 concomitant with the proton transfer from H211 to the hydroxylic group. However, it was impossible to locate **I8** where the H211 residue assisted as a base. H211 residue plays a distinctive electrostatic/structural role that stands out from the covalent participation that it usually plays in other systems.[38,40] Furthermore, we can observe that the interaction of H211 with the reactive water occurs by means of a bridge of two water molecules. As a matter of fact, this arrangement allows the reactive water to keep its proton oriented toward the Sγ:C244 atom of the cysteine (see **I5** snapshot in Figure 5).

Analysis of the active site in the **I5** intermediate (see **I5** in Figure 5) allows us to propose the direct proton transfer from the water molecule to the Sγ:C244 atom and the attack of the hydroxylic group to the C1:Pep atom. The resulting free energy surface indicates that this reaction proceeds in a concerted manner with a free energy barrier associated with the transition state $\mathbf{TS}^{\mathbf{I5-P1}}$ of 23.5 kcal·mol$^{-1}$. The atomic charge analysis (see Table S3 in the SI) reveals a late proton transfer in the transition state, while the attack of the oxygen atom of the water molecule on the C1:Pep atom is found to be advanced and the charge on the Sγ:C244 atom becomes slightly more negative at the transition state $\mathbf{TS}^{\mathbf{I5-P1}}$. Then, the deacylation step will take place through this reaction mechanism. As in the case of the acylation step, the rate limiting $\mathbf{TS}^{\mathbf{I5-P1}}$ was optimized at the PBE+D3(BJ)/MM level (see Figure 6). The comparison between the structures localized and characterized using a high level method (PBE +D3(BJ)/MM) and low level method (PM6/MM) shows structural similarities giving rise to the robustness of our conclusions based on DFT/MM corrections over QM/MM MD simulations using the PM6/MM method. A structural analysis of the stationary structures in the deacylation stage reveals the presence of two water molecules in the position of the first fragment of the peptide released where the H211 residue disposes them in a reactive orientation.

The deacylation step shows a free energy barrier of 25.2 kcal·mol$^{-1}$ from **I5**, which could be dictating the rate limiting step of the full process. Neverthess, considering the energies of this $\mathbf{TS}^{\mathbf{I5-P1}}$ and the first transition state of the acylation process, $\mathbf{TS}^{\mathbf{R1\text{-}I4}}$, relative to the reactants' state (23.4 and 23.5 kcal·mol$^{-1}$, respectively, as shown in Figure 4), together with the intrinsic uncertainty of the employed computational method, both steps would be contributing to determine the kinetics of the overall reaction. The relative energy of the product of the reaction, P1, describes the reaction as an exergonic process in thermodynamic terms (−14.6 kcal·mol$^{-1}$). The deacylation process occurs similarly to some of those studies done in our group and their energies are quite comparable,[38,40] 26.6 kcal·mol$^{-1}$ and 22.8 kcal·mol$^{-1}$, respectively. Nevertheless, due to the different distribution of the active site, the role that plays the catalytic H211 residue is significantly different.

## 5. CONCLUSIONS

In this study, the reaction mechanism of proteolysis of the RgpB gingipain was revealed at the atomic level by means of QM/MM MD simulations. Initially, due to the lack of data of the protonation states of the Cys/His catalytic dyad, their roles in the reaction mechanism, and their particular arrangement in the active site, three structures with different protonation states (**R1**, **R2**, and **R3**) were explored by performing classical MD simulations. From the analysis of the MD simulations, conformational changes of a flexible loop that affects the interactions between the peptide and E152 and H211 residues were observed. While **R2** and **R3** show a wide range of conformations or an open nonreactive conformation of the flexible loop, **R1** is always in a closed and reactive conformation keeping the interactions between E152 and H211 residues and the peptide.

Starting from **R1**, an exhaustive mechanistic study has been performed by the exploration of all of the plausible reaction mechanisms catalyzed by RgpB gingipain. The most likely one consists of two steps for the acylation stage and one step for the deacylation stage. In the first step of the acylation, the nucleoplilic attack of the Sγ:C244 atom to the C1 atom of the peptide and the proton transfer from the Sγ:C244 atom to the N1 atom of the peptide takes place in a concerted manner. Subsequently, the breaking of the peptide bond takes place with the localization and characterization of an I5 intermediate both structurally and energetically similar to the one reported in the study of the protease mechanism of Human Legumain carried out by Elsässer and co-workers.[39] The rate limiting step for the acylation is the first step with an activation free energy of 23.4 kcal·mol$^{-1}$, very close to the value that can be deduced from previous kinetic reports, 22.8 kcal·mol$^{-1}$.[21] Finally, the deacylation stage takes place in a single step where a water molecule attacks the C1 of the peptide and one proton of the water is transferred to the Sγ:C244 atom in a concerted way with a free energy barrier of 23.5 kcal·mol$^{-1}$ relative to the R1 reactant state.

It is noteworthy that along this reaction mechanism, promoted by the unusual arrangement of the catalytic dyad in the active site, H211 residue does not play a role as a base. Alternatively, it plays an essential role in the orientation of the two water molecules located in the active site after the first fragment of the peptide is release in the acylation step. This reaction mechanism also suggests the crucial role played by the substrate in promoting enzyme reactivity. This role should be taken into account for future drug designing of inhibitors for RgpB for the treatment of Alzheimer's disease and other human conditions.

## ■ ASSOCIATED CONTENT

### Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jcim.1c00666.

> Details to perform the PMFs; standard deviations of RMSD values for **R1**, **R2**, and **R3**; snapshot of the active site of RgpB protease in the R1 state; all the PMFs computed; free energy profile of the alternative reaction mechanisms; main average interaction energies between the peptide and the protein residues; CHelpG charges of the key atoms calculated on the localized PBE+D3(BJ)/MM stationary points; quantum region TS Cartesian coordinates computed using PBE-D3(BJ)/MM (PDF)

## ■ AUTHOR INFORMATION

### Corresponding Authors

**Maite Roca** − *BioComp Group, Institute of Advanced Materials (INAM), Universidad Jaume I, 12071 Castellón, Spain;* orcid.org/0000-0003-0937-4722;
Email: mroca@uji.es

**Vicent Moliner** − *BioComp Group, Institute of Advanced Materials (INAM), Universidad Jaume I, 12071 Castellón, Spain*; orcid.org/0000-0002-3665-3391;
Email: moliner@uji.es

## Authors

**Santiago Movilla** − *BioComp Group, Institute of Advanced Materials (INAM), Universidad Jaume I, 12071 Castellón, Spain*

**Sergio Martí** − *BioComp Group, Institute of Advanced Materials (INAM), Universidad Jaume I, 12071 Castellón, Spain*; orcid.org/0000-0002-1087-7143

Complete contact information is available at:
https://pubs.acs.org/10.1021/acs.jcim.1c00666

## Notes

The authors declare no competing financial interest.

## ■ ACKNOWLEDGMENTS

We would like to thank the Spanish Ministerio de Ciencia e Innovación (grant PGC2018-094852−B-C21), Generalitat Valenciana (grant AICO/2019/195), and Universitat Jaume I (grant UJI-B2020-03 and UJI-B2019-43). Santiago Movilla and Sergio Martí thank the Generalitat Valenciana for a Grisolia Ph.D. grant (GRISOLIAP/2019/064). The authors acknowledge computational resources from the Servei d'Informàtica of Universitat Jaume I.

## ■ REFERENCES

(1) Sharma, P.; Srivastava, P.; Seth, A.; Tripathi, P. N.; Banerjee, A. G.; Shrivastava, S. K. Comprehensive Review of Mechanisms of Pathogenesis Involved in Alzheimer's Disease and Potential Therapeutic Strategies. *Prog. Neurobiol.* **2019**, *174*, 53−89.

(2) De Strooper, B.; Karran, E. The Cellular Phase of Alzheimer's Disease. *Cell* **2016**, *164*, 603−615.

(3) Mirakhur, A.; Craig, D.; Hart, D. J.; Mcllroy, S. P.; Passmore, A. P. Behavioural and Psychological Syndromes in Alzheimer's Disease. *Int. J. Geriatr. Psychiatry* **2004**, *19*, 1035−1039.

(4) Kumar, A.; Singh, A.; Ekavali. A Review on Alzheimer's Disease Pathophysiology and Its Management: An Update. *Pharmacol. Rep.* **2015**, *67*, 195−203.

(5) Huang, Y.; Mucke, L. Alzheimer Mechanisms and Therapeutic Strategies. *Cell* **2012**, *148*, 1204−1222.

(6) Kaye, E. K.; Valencia, A.; Baba, N.; Spiro, A.; Dietrich, T.; Garcia, R. I. Tooth Loss and Periodontal Disease Predict Poor Cognitive Function in Older Men. *J. Am. Geriatr. Soc.* **2010**, *58*, 713−718.

(7) Gatz, M.; Mortimer, J. A.; Fratiglioni, L.; Johansson, B.; Berg, S.; Reynolds, C. A.; Pedersen, N. L. Potentially Modifiable Risk Factors for Dementia in Identical Twins. *Alzheimer's Dementia* **2006**, *2*, 110−117.

(8) Stein, P. S.; Desrosiers, M.; Donegan, S. J.; Yepes, J. F.; Kryscio, R. J. Tooth Loss, Dementia and Neuropathology in the Nun Study. *J. Am. Dent. Assoc., JADA* **2007**, *138*, 1314−1322.

(9) Kamer, A. R.; Pirraglia, E.; Tsui, W.; Rusinek, H.; Vallabhajosula, S.; Mosconi, L.; Yi, L.; McHugh, P.; Craig, R. G.; Svetcov, S.; Linker, R.; Shi, C.; Glodzik, L.; Williams, S.; Corby, P.; Saxena, D.; de Leon, M. J. Periodontal Disease Associates with Higher Brain Amyloid Load in Normal Elderly. *Neurobiol. Aging* **2015**, *36*, 627−633.

(10) Noble, J. M.; Borrell, L. N.; Papapanou, P. N.; Elkind, M. S. V.; Scarmeas, N.; Wright, C. B. Periodontitis Is Associated with Cognitive Impairment among Older Adults: Analysis of NHANES-III. *J. Neurol., Neurosurg. Psychiatry* **2009**, *80*, 1206−1211.

(11) Dominy, S. S.; Lynch, C.; Ermini, F.; Benedyk, M.; Marczyk, A.; Konradi, A.; Nguyen, M.; Haditsch, U.; Raha, D.; Griffin, C.; Holsinger, L. J.; Arastu-Kapur, S.; Kaba, S.; Lee, A.; Ryder, M. I.; Potempa, B.; Mydel, P.; Hellvard, A.; Adamowicz, K.; Hasturk, H.; Walker, G. D.; Reynolds, E. C.; Faull, R. L. M.; Curtis, M. A.; Dragunow, M.; Potempa, J. Porphyromonas Gingivalis in Alzheimer's Disease Brains: Evidence for Disease Causation and Treatment with Small-Molecule Inhibitors. *Sci. Adv.* **2019**, *5*, No. eaau3333.

(12) Mottram, J. C.; Helms, M. J.; Coombs, G. H.; Sajid, M. Clan CD Cysteine Peptidases of Parasitic Protozoa. *Trends Parasitol.* **2003**, *19*, 182−187.

(13) Mayrand, D.; Holt, S. C. Biology of Asaccharolytic Black-Pigmented Bacteroides Species. *Microbiol. Rev.* **1988**, *52*, 134−152.

(14) Mahendra, J.; Mahendra, L.; Kurian, V. M.; Jaishankar, K.; Mythilli, R. Prevalence of Periodontal Pathogens in Coronary Atherosclerotic Plaque of Patients Undergoing Coronary Artery Bypass Graft Surgery. *J. Maxillofac. Oral Surg.* **2009**, *8*, 108−113.

(15) Katz, J.; Chegini, N.; Shiverick, K. T.; Lamont, R. J. Localization of P. Gingivalis in Preterm Delivery Placenta. *J. Dent. Res.* **2009**, *88*, 575−578.

(16) Ishikawa, M.; Yoshida, K.; Okamura, H.; Ochiai, K.; Takamura, H.; Fujiwara, N.; Ozaki, K. Oral Porphyromonas Gingivalis Translocates to the Liver and Regulates Hepatic Glycogen Synthesis through the Akt/GSK-3$\beta$ Signaling Pathway. *Biochim. Biophys. Acta, Mol. Basis Dis.* **2013**, *1832*, 2035−2043.

(17) Mougeot, J.-L. C.; Stevens, C. B.; Paster, B. J.; Brennan, M. T.; Lockhart, P. B.; Mougeot, F. K. B. Porphyromonas Gingivalis Is the Most Abundant Species Detected in Coronary and Femoral Arteries. *J. Oral Microbiol.* **2017**, *9*, 1281562.

(18) Potempa, J.; Travis, J. Porphyromonas Gingivalis Proteinases in Periodontitis, a Review. *Acta Biochim. Polym.* **1996**, *43*, 455−465.

(19) Potempa, J.; Pavloff, N.; Travis, J. Porphyromonas Gingivalis: A Proteinase/Gene Accounting Audit. *Trends in Microbiology*; Elsevier Current Trends, 1995; pp 430−434.

(20) Curtis, M. A.; Kuramitsu, H. K.; Lantz, M.; Macrina, F. L.; Nakayama, K.; Potempa, J.; Reynolds, E. C.; Aduse-Opoku, J. Molecular Genetics and Nomenclature of Proteases of Porphyromonas Gingivalis. *J. Periodontal Res.* **1999**, *34*, 464−472.

(21) Ally, N.; Whisstock, J. C.; Sieprawska-Lupa, M.; Potempa, J.; Le Bonniec, B. F.; Travis, J.; Pike, R. N. Characterization of the Specificity of Arginine-Specific Gingipains from Porphyromonas Gingivalis Reveals Active Site Differences between Different Forms of the Enzymes. *Biochemistry* **2003**, *42*, 11693−11700.

(22) Eichinger, A.; Beisel, H. G.; Jacob, U.; Huber, R.; Medrano, F. J.; Banbula, A.; Potempa, J.; Travis, J.; Bode, W. Crystal Structure of Gingipain R: An Arg-Specific Bacterial Cysteine Proteinase with a Caspase-like Fold. *EMBO J.* **1999**, *18*, 5453−5462.

(23) Potempa, J.; Mikolajczyk-Pawlinska, J.; Brassell, D.; Nelson, D.; Thøgersen, I. B.; Enghild, J. J.; Travis, J. Comparative Properties of Two Cysteine Proteinases (Gingipains R), the Products of Two Related but Individual Genes of Porphyromonas Gingivalis. *J. Biol. Chem.* **1998**, *273*, 21648−21657.

(24) Bedi, G. S. Purification and Characterization of Lysine- and Arginine-Specific Gingivain Proteases from Porphyromonas Gingivalis. *Prep. Biochem.* **1994**, *24*, 251−261.

(25) Fujimura, S.; Shibata, Y.; Nakamura, T. Purification and Partial Characterization of a Lysine-Specific Protease of Porphyromonas Gingivalis. *FEMS Microbiol. Lett.* **1993**, *113*, 133−137.

(26) Hinode, D.; Hayashi, H.; Nakamura, R. Purification and Characterization of Three Types of Proteases from Culture Supernatants of Porphyromonas Gingivalis. *Infect. Immun.* **1991**, *59*, 3060−3068.

(27) Bleeg, H. S. Non-Specific Cleavage of Collagen by Proteinases in the Presence of Sodium Dodecyl Sulfate. *Eur. J. Oral Sci.* **1990**, *98*, 235−241.

(28) Griffen, A. L.; Becker, M. R.; Lyons, S. R.; Moeschberger, M. L.; Leys, E. J. Prevalence of Porphyromonas Gingivalisand Periodontal Health Status. *J. Clin. Microbiol.* **1998**, *36*, 3239−3242.

(29) Harrison, M. J.; Burton, N. A.; Hillier, I. H. Catalytic Mechanism of the Enzyme Papain: Predictions with a Hybrid Quantum Mechanical/Molecular Mechanical Potential. *J. Am. Chem. Soc.* **1997**, *119*, 12285−12291.

(30) Wei, D.; Huang, X.; Liu, J.; Tang, M.; Zhan, C. G. Reaction Pathway and Free Energy Profile for Papain-Catalyzed Hydrolysis of N-Acetyl-Phe-Gly 4-Nitroanilide. *Biochemistry* **2013**, *52*, 5145−5154.

(31) Arad, D.; Langridge, R.; Kollman, P. A. A Simulation of the Sulfur Attack in the Catalytic Pathway of Papain Using Molecular Mechanics and Semiempirical Quantum Mechanics. *J. Am. Chem. Soc.* **1990**, *112*, 491−502.

(32) Welsh, W. J.; Lin, Y. Discussion of the Catalytic Pathway of Cysteine Proteases Based on AM1 Calculations. *J. Mol. Struct.: THEOCHEM* **1997**, *401*, 315−326.

(33) Sulpizi, M.; Rothlisberger, U.; Carloni, P. Molecular Dynamics Studies of Caspase-3. *Biophys. J.* **2003**, *84*, 2207−2215.

(34) Sulpizi, M.; Laio, A.; VandeVondele, J.; Cattaneo, A.; Rothlisberger, U.; Carloni, P. Reaction Mechanism of Caspases: Insights from QM/MM Car-Parrinello Simulations. *Proteins: Struct., Funct., Genet.* **2003**, *52*, 212−224.

(35) Miscione, G. P.; Calvaresi, M.; Bottoni, A. Computational Evidence for the Catalytic Mechanism of Caspase-7. A DFT Investigation. *J. Phys. Chem. B* **2010**, *114*, 4637−4645.

(36) Angelides, K. J.; Fink, A. L. Mechanism of Thiol Protease Catalysis: Detection and Stabilization of a Tetrahedral Intermediate in Papain Catalysis. *Biochemistry* **1979**, *18*, 2363−2369.

(37) Ŝtrajbl, M.; Florián, J.; Warshel, A. Ab Initio Evaluation of the Free Energy Surfaces for the General Base/Acid Catalyzed Thiolysis of Formamide and the Hydrolysis of Methyl Thiolformate: A Reference Solution Reaction for Studies of Cysteine Proteases. *J. Phys. Chem. B* **2001**, *105*, 4471−4484.

(38) Arafet, K.; Ferrer, S.; Moliner, V. Computational Study of the Catalytic Mechanism of the Cruzain Cysteine Protease. *ACS Catal.* **2017**, *7*, 1207−1215.

(39) Elsässer, B.; Zauner, F. B.; Messner, J.; Soh, W. T.; Dall, E.; Brandstetter, H. 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.* **2017**, *7*, 5585−5593.

(40) Ŝwiderek, K.; Moliner, V. Revealing the Molecular Mechanisms of Proteolysis of SARS-CoV-2 Mproby QM/MM Computational Methods. *Chem. Sci.* **2020**, *11*, 10626−10630.

(41) Ramos-Guzmán, C. A.; Ruiz-Pernía, J. J.; Tuñón, I. Unraveling the SARS-CoV-2 Main Protease Mechanism Using Multiscale Methods. *ACS Catal.* **2020**, *10*, 12544−12554.

(42) Ramos-Guzmán, C. A.; Zinovjev, K.; Tuñón, I. Modeling Caspase-1 Inhibition: Implications for Catalytic Mechanism and Drug Design. *Eur. J. Med. Chem.* **2019**, *169*, 159−167.

(43) Arafet, K.; Ferrer, S.; González, F. V.; Moliner, V. Quantum Mechanics/Molecular Mechanics Studies of the Mechanism of Cysteine Protease Inhibition by Peptidyl-2,3-Epoxyketones. *Phys. Chem. Chem. Phys.* **2017**, *19*, 12740−12748.

(44) Arafet, K.; Ŝwiderek, K.; Moliner, V. Computational Study of the Michaelis Complex Formation and the Effect on the Reaction Mechanism of Cruzain Cysteine Protease. *ACS Omega* **2018**, *3*, 18613−18622.

(45) Arafet, K.; Serrano-Aparicio, N.; Lodola, A.; Mulholland, A. J.; González, F. V.; Ŝwiderek, K.; Moliner, V. Mechanism of Inhibition of SARS-CoV-2 M pro by N3 Peptidyl Michael Acceptor Explained by QM/MM Simulations and Design of New Derivatives with Tunable Chemical Reactivity. *Chem. Sci.* **2021**, *12*, 1433−1444.

(46) Hanada, K.; Tamai, M.; Yamagishi, M.; Ohmura, S.; Sawada, J.; Tanaka, I. Isolation and Characterization of E−64, a New Thiol Protease Inhibitor. *Agric. Biol. Chem.* **1978**, *42*, 523−528.

(47) Rangarajan, M.; Smith, S. J. M.; U, S.; Curtis, M. A. Biochemical Characterization of the Arginine-Specific Proteases of Porphyromonas Gingivalis W50 Suggests a Common Precursor. *Biochem. J.* **1997**, *323*, 701−709.

(48) Bateman, A.; Martin, M.-J.; Orchard, S.; Magrane, M.; Agivetova, R.; Ahmad, S.; Alpi, E.; Bowler-Barnett, E. H.; Britto, R.; Bursteinas, B.; Bye-A-Jee, H.; Coetzee, R.; Cukura, A.; Da Silva, A.; Denny, P.; Dogan, T.; Ebenezer, T.; Fan, J.; Castro, L. G.; Garmiri, P.; Georghiou, G.; Gonzales, L.; Hatton-Ellis, E.; Hussein, A.; Ignatchenko, A.; Insana, G.; Ishtiaq, R.; Jokinen, P.; Joshi, V.; Jyothi, D.; Lock, A.; Lopez, R.; Luciani, A.; Luo, J.; Lussi, Y.; MacDougall, A.; Madeira, F.; Mahmoudy, M.; Menchi, M.; Mishra, A.; Moulang, K.; Nightingale, A.; Oliveira, C. S.; Pundir, S.; Qi, G.; Raj, S.; Rice, D.; Lopez, M. R.; Saidi, R.; Sampson, J.; Sawford, T.; Speretta, E.; Turner, E.; Tyagi, N.; Vasudev, P.; Volynkin, V.; Warner, K.; Watkins, X.; Zaru, R.; Zellner, H.; Bridge, A.; Poux, S.; Redaschi, N.; Aimo, L.; Argoud-Puy, G.; Auchincloss, A.; Axelsen, K.; Bansal, P.; Baratin, D.; Blatter, M.-C.; Bolleman, J.; Boutet, E.; Breuza, L.; Casals-Casas, C.; de Castro, E.; Echioukh, K. C.; Coudert, E.; Cuche, B.; Doche, M.; Dornevil, D.; Estreicher, A.; Famiglietti, M. L.; Feuermann, M.; Gasteiger, E.; Gehant, S.; Gerritsen, V.; Gos, A.; Gruaz-Gumowski, N.; Hinz, U.; Hulo, C.; Hyka-Nouspikel, N.; Jungo, F.; Keller, G.; Kerhornou, A.; Lara, V.; Le Mercier, P.; Lieberherr, D.; Lombardot, T.; Martin, X.; Masson, P.; Morgat, A.; Neto, T. B.; Paesano, S.; Pedruzzi, I.; Pilbout, S.; Pourcel, L.; Pozzato, M.; Pruess, M.; Rivoire, C.; Sigrist, C.; Sonesson, K.; Stutz, A.; Sundaram, S.; Tognolli, M.; Verbregue, L.; Wu, C. H.; Arighi, C. N.; Arminski, L.; Chen, C.; Chen, Y.; Garavelli, J. S.; Huang, H.; Laiho, K.; McGarvey, P.; Natale, D. A.; Ross, K.; Vinayaka, C. R.; Wang, Q.; Wang, Y.; Yeh, L.-S.; Zhang, J.; Ruch, P.; Teodoro, D. UniProt: The Universal Protein Knowledgebase in 2021. *Nucleic Acids Res.* **2021**, *49*, D480−D489.

(49) Olsson, M. H. M.; Søndergaard, C. R.; Rostkowski, M.; Jensen, J. H. PROPKA3: Consistent Treatment of Internal and Surface Residues in Empirical p K a Predictions. *J. Chem. Theory Comput.* **2011**, *7*, 525−537.

(50) Li, P.; Song, L. F.; Merz, K. M. Parameterization of Highly Charged Metal Ions Using the 12−6-4 LJ-Type Nonbonded Model in Explicit Water. *J. Phys. Chem. B* **2015**, *119*, 883−895.

(51) 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.* **1983**, *79*, 926−935.

(52) Salomon-Ferrer, R.; Case, D. A.; Walker, R. C. An Overview of the Amber Biomolecular Simulation Package. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2013**, *3*, 198−210.

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

(54) Stewart, J. J. P. Optimization of Parameters for Semiempirical Methods V: Modification of NDDO Approximations and Application to 70 Elements. *J. Mol. Model.* **2007**, *13*, 1173−1213.

(55) Götz, A. W.; Williamson, M. J.; Xu, D.; Poole, D.; Le Grand, S.; Walker, R. C. Routine Microsecond Molecular Dynamics Simulations with AMBER on GPUs. 1. Generalized Born. *J. Chem. Theory Comput.* **2012**, *8*, 1542−1555.

(56) 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.* **2013**, *9*, 3878−3888.

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

(58) Darden, T.; York, D.; Pedersen, L. Particle Mesh Ewald: An N · log(N) Method for Ewald Sums in Large Systems. *J. Chem. Phys.* **1993**, *98*, 10089−10092.

(59) Uberuaga, B. P.; Anghel, M.; Voter, A. F. Synchronization of Trajectories in Canonical Molecular-Dynamics Simulations: Observation, Explanation, and Exploitation. *J. Chem. Phys.* **2004**, *120*, 6363−6374.

(60) Sindhikara, D. J.; Kim, S.; Voter, A. F.; Roitberg, A. E. Bad Seeds Sprout Perilous Dynamics: Stochastic Thermostat Induced Trajectory Synchronization in Biomolecules. *J. Chem. Theory Comput.* **2009**, *5*, 1624−1631.

(61) Ryckaert, J.-P.; Ciccotti, G.; Berendsen, H. J. Numerical Integration of the Cartesian Equations of Motion of a System with Constraints: Molecular Dynamics of n-Alkanes. *J. Comput. Phys.* **1977**, *23*, 327−341.

(62) Kräutler, V.; van Gunsteren, W. F.; Hünenberger, P. H. A Fast SHAKE Algorithm to Solve Distance Constraint Equations for Small Molecules in Molecular Dynamics Simulations. *J. Comput. Chem.* **2001**, *22*, 501−508.

(63) Verlet, L. Computer "Experiments" on Classical Fluids. I. Thermodynamical Properties of Lennard-Jones Molecules. *Phys. Rev.* **1967**, *159*, 98−103.

(64) Roe, D. R.; Cheatham, T. E. PTRAJ and CPPTRAJ: Software for Processing and Analysis of Molecular Dynamics Trajectory Data. *J. Chem. Theory Comput.* **2013**, *9*, 3084−3095.

(65) Roux, B. The Calculation of the Potential of Mean Force Using Computer Simulations. *Comput. Phys. Commun.* **1995**, *91*, 275−282.

(66) Kumar, S.; Rosenberg, J. M.; Bouzida, D.; Swendsen, R. H.; Kollman, P. A. Multidimensional Free-Energy Calculations Using the Weighted Histogram Analysis Method. *J. Comput. Chem.* **1995**, *16*, 1339−1350.

(67) Torrie, G. M.; Valleau, J. P. Nonphysical Sampling Distributions in Monte Carlo Free-Energy Estimation: Umbrella Sampling. *J. Comput. Phys.* **1977**, *23*, 187−199.

(68) Kästner, J.; Thiel, W. Bridging the Gap between Thermodynamic Integration and Umbrella Sampling Provides a Novel Analysis Method: "Umbrella Integration. *J. Chem. Phys.* **2005**, *123*, 144104.

(69) Martí, S. QMCube (QM3): An All-Purpose Suite for Multiscale QM/MM Calculations. *J. Comput. Chem.* **2021**, *42*, 447−457.

(70) Rodriguez, A.; Laio, A. Clustering by Fast Search and Find of Density Peaks. *Science (Washington, DC, U. S.)* **2014**, *344*, 1492−1496.

(71) Perdew, J. P.; Burke, K.; Ernzerhof, M. Generalized Gradient Approximation Made Simple. *Phys. Rev. Lett.* **1996**, *77*, 3865−3868.

(72) Ditchfield, R.; Hehre, W. J.; Pople, J. A. Self□Consistent Molecular□Orbital Methods. IX. An Extended Gaussian□Type Basis for Molecular□Orbital Studies of Organic Molecules. *J. Chem. Phys.* **1971**, *54*, 724−728.

(73) Hehre, W. J.; Ditchfield, R.; Pople, J. A. Self—Consistent Molecular Orbital Methods. XII. Further Extensions of Gaussian—Type Basis Sets for Use in Molecular Orbital Studies of Organic Molecules. *J. Chem. Phys.* **1972**, *56*, 2257−2261.

(74) Clark, T.; Chandrasekhar, J.; Spitznagel, G. W.; Schleyer, P. V. R. Efficient Diffuse Function-Augmented Basis Sets for Anion Calculations. III. The 3-21+G Basis Set for First-Row Elements, Li-F. *J. Comput. Chem.* **1983**, *4*, 294−301.

(75) Grimme, S.; Ehrlich, S.; Goerigk, L. Effect of the Damping Function in Dispersion Corrected Density Functional Theory. *J. Comput. Chem.* **2011**, *32*, 1456−1465.

(76) Breneman, C. M.; Wiberg, K. B. Determining Atom-Centered Monopoles from Molecular Electrostatic Potentials. The Need for High Sampling Density in Formamide Conformational Analysis. *J. Comput. Chem.* **1990**, *11*, 361−373.

(77) Case, D. A.; Darden, T.; Cheatham, T. E., III; Simmerling, C.; Roitberg, A.; Wang, J.; Götz, A. W.; Kolossváry, I.; Paesani, F.; Liu, J.; Wu, X.; Steinbrecher, T.; et al. *Amber 14 Reference Manual Principal Contributors to the Current Codes*; University of California: San Francisco, 2014.

(78) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Petersson, G. A.; Nakatsuji, H.; Li, X.; Caricato, M.; Marenich, A.; Bloino, J.; Janesko, B. G.; Gomperts, R.; Mennucci, B.; Hratchian, H. P.; Ortiz, J. V.; Izmaylov, A. F.; Sonnenberg, J. L.; Williams-Young, D.; Ding, F.; Lipparini, F.; Egidi, F.; Goings, J.; Peng, B.; Petrone, A.; Henderson, T.; Ranasinghe, D.; Zakrzewski, V. G.; Gao, J.; Rega, N.; Zheng, G.; Liang, W.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Vreven, T.; Throssell, K.; J. A. Montgomery, J.; Peralta, J. E.; Ogliaro, F.; Bearpark, M.; Heyd, J. J.; Brothers, E.; Kudin, K. N.; Staroverov, V. N.; Keith, T.; Kobayashi, R.; Normand, J.; Raghavachari, K.; Rendell, A.; Burant, J. C.; Iyengar, S. S.; Tomasi, J.; Cossi, M.; Millam, J. M.; Klene, M.; Adamo, C.; Cammi, R.; Ochterski, J. W.; Martin, R. L.; Morokuma, K.; Farkas, O.; Foresman, J. B.; Fox, D. J. *Gaussian 09*, revision A.02; Gaussian, Inc.: Wallingford, CT, 2009.

(79) Field, M. J.; Albe, M.; Bret, C.; Proust-De Martin, F.; Thomas, A. The Dynamo Library for Molecular Simulations Using Hybrid Quantum Mechanical and Molecular Mechanical Potentials. *J. Comput. Chem.* **2000**, *21*, 1088−1100.

(80) Field, M. J. *A Practical Introduction to the Simulation of Molecular Systems*; Cambridge University Press, 1999.

(81) Mazmanian, K.; Sargsyan, K.; Grauffel, C.; Dudev, T.; Lim, C. Preferred Hydrogen-Bonding Partners of Cysteine: Implications for Regulating Cys Functions. *J. Phys. Chem. B* **2016**, *120*, 10288−10296.

(82) Dall'Acqua, W.; Carter, P. Substrate-Assisted Catalysis: Molecular Basis and Biological Significance. *Protein Sci.* **2000**, *9*, 1−9.

# Supporting Information

# Unrevealing the Proteolytic Activity of RgpB Gingipain from Computational Simulations

Santiago Movilla, Sergio Martí, Maite Roca* and Vicent Moliner*
Institute of Advanced Materials (INAM), Universidad Jaume I, 12071, Castellón, Spain

*correspondence should be addressed to:
Maite Roca: mroca@uji.es
Vicent Moliner: moliner@uji.es

**Table S1.** Detailed description of the collective variables (CV, defined as ξi) used to computed the PMF calculations in the study of the reaction catalyzed by the RgpB gingipain enzyme. Collective variable intervals (Δξ) are expressed in Å.

| **Step** | **ξ** | **CV** | **Δξ(Å)** | **Number of Windows** |
|---|---|---|---|---|
| R1-I1 | ξ1 | [Hγ:Cys244-Sγ:Cys244]-[Hγ:Cys211-O1:Pep] | 0.1 | 30 |
| | ξ2 | [C1:Pep-Sγ:Cys244] | 0.05 | 35 |
| R1-I4 | ξ3 | [Hγ:Cys244-Sγ:Cys244]-[Hγ:Cys211-N1:Pep] | 0.1 | 33 |
| | ξ4 | [C1:Pep-Sγ:Cys244] | 0.05 | 35 |
| I1-I4 | ξ5 | [Hγ:Cys244-O1:Pep] | 0.05 | 32 |
| | ξ6 | [Hγ:Cys244-N1:Pep] | 0.05 | 32 |
| I5-I6 | ξ7 | [Hw:Wat-Ow:Wat]-[Hw:Wat-O1:Pep] | 0.1 | 30 |
| | ξ8 | [Ow-:Wat-C1:Pep] | 0.05 | 34 |
| I5-P1 | ξ9 | [C1:Pep-Sγ:Cys244]-[C1:Pep-Ow:Wat] | 0.1 | 30 |
| | ξ10 | [Hw:Wat-Ow:Wat]-[Hw:Wat-Sγ:Pep] | 0.1 | 30 |
| I6-P1 | ξ11 | [Hw:Wat-O1:Pep]-[Hw:Wat-Sγ:Cys244] | 0.1 | 33 |
| | ξ12 | [C1:Pep-Sγ:Cys244] | 0.05 | 34 |
| I7-P2 | ξ13 | [Hw:Wat-O1:Pep]-[Hw:Wat-Sγ:Cys244] | 0.1 | 33 |
| | ξ14 | [C1:Pep-Sγ:Cys244] | 0.05 | 34 |
| I1-I2 | ξ15 | [O1:Pep-Hγ:Cys244]-[Nδ:His211-Hγ:Cys244] | 0.05 | 51 |
| I2-I4 | ξ16 | [Nδ:His211-Hγ:Cys244]-[N1:Pep-Hγ:Cys244] | 0.05 | 54 |
| I4-I5 | ξ17 | [C1:Pep-N1:Pep] | 0.05 | 76 |
| I6-I7 | ξ18 | [Hw:Wat-Ow:Wat]-[Hw:Wat-Nδ:His211] | 0.05 | 76 |

**Table S2.** Relative standard deviations of RMSD values calculated for the backbone atoms of the protein and on the atoms of the active site along the classical equilibrium trajectories of **R1**, **R2** and **R3**.

| | **Backbone** | **Active Site** |
|---|---|---|
| **R1** | 8% | 12% |
| **R2** | 25% | 31% |
| **R3** | 16% | 51% |

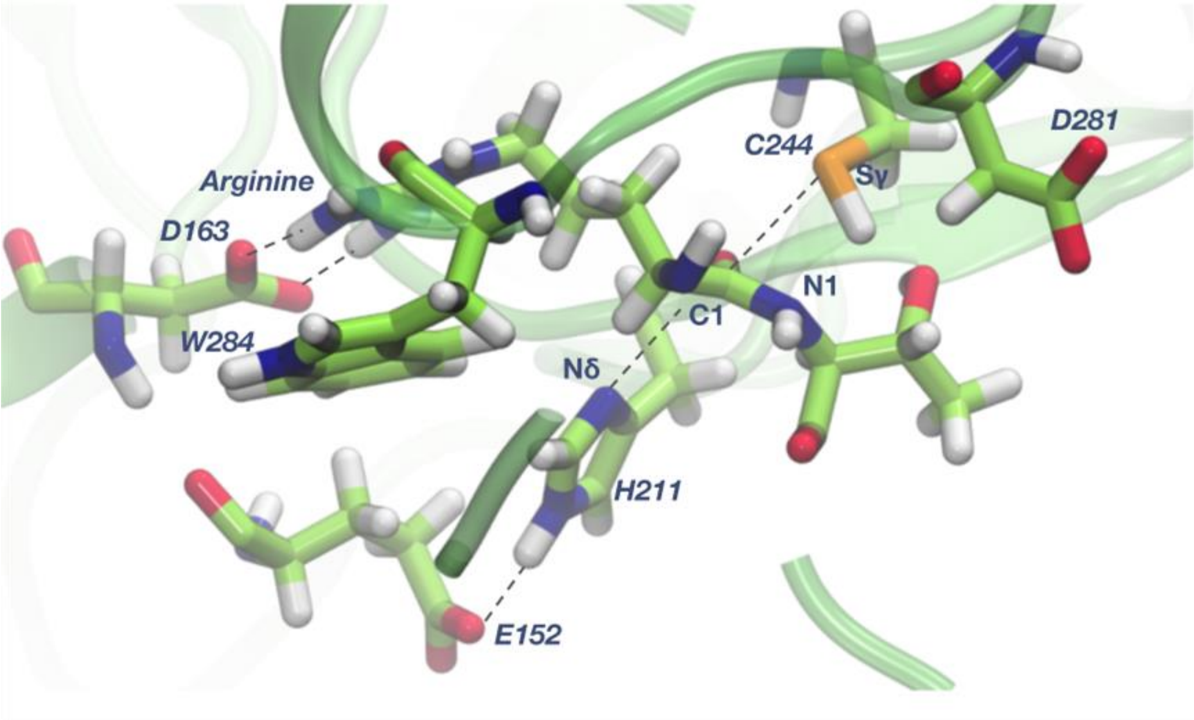


**Figure S1.** Structure of the active site of RgpB protease of R1 reactant state.

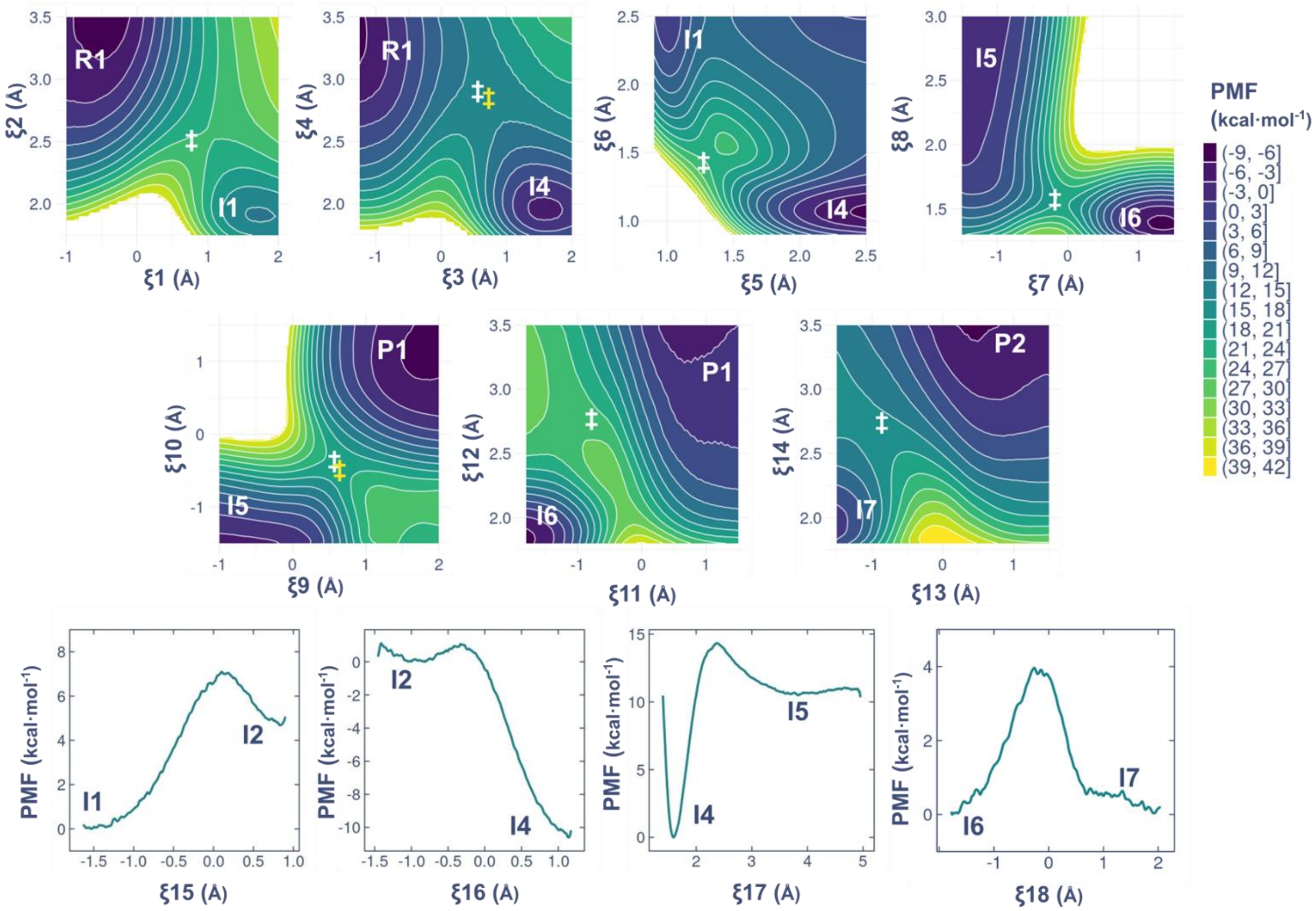


**Figure S2.** Free energy surfaces computed at PM6/MM level for the reaction mechanisms considered for the proteolysis catalyzed by RgpB gingipain. The values of energies are in kcal·mol$^{-1}$. Each level in the contour plots represent a 3 kcal·mol$^{-1}$ variation. A complete description of the collective variables and the PMF parameters are shown in table S1. White and yellow double daggers point the coordinates of the localized PM6/MM and PBE-D3(BJ)/MM transitions states, respectively.

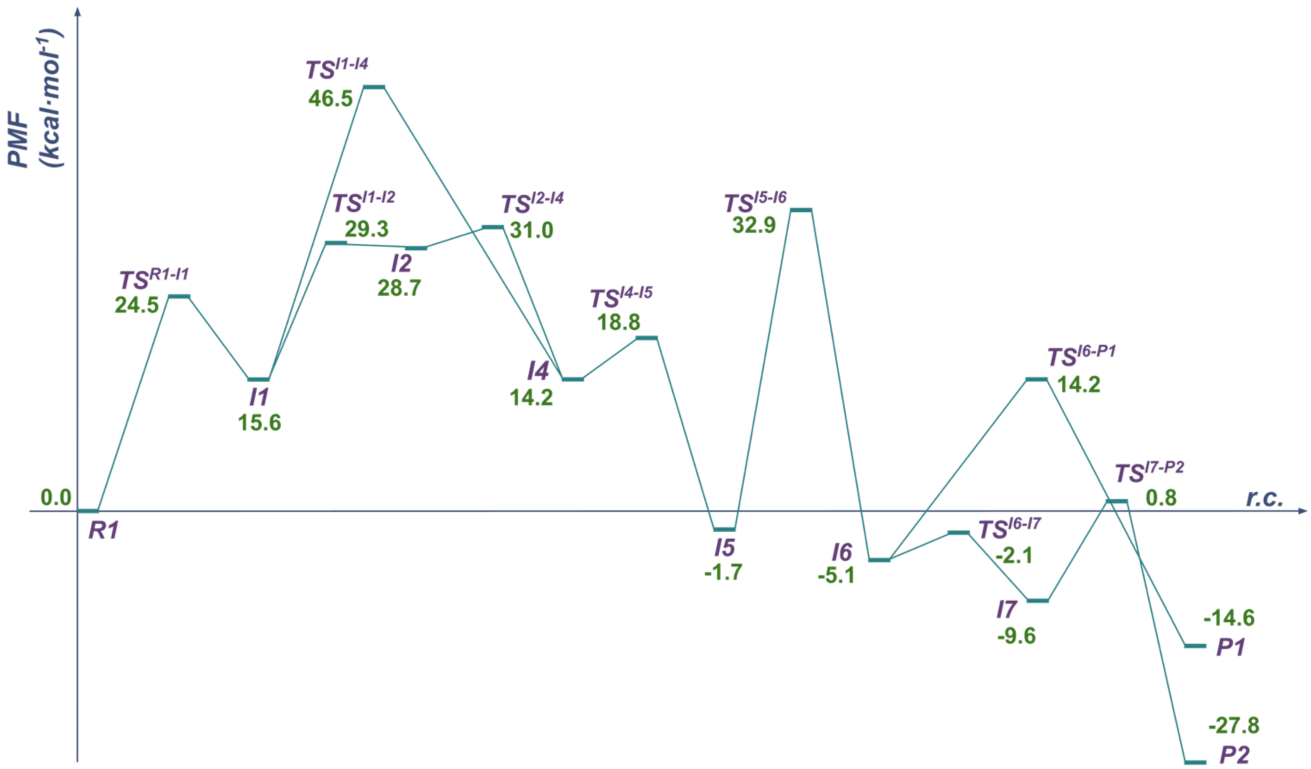


**Figure S3.** PBE+D3(BJ):PM6/MM free energy profile of the alternative reaction pathways computed for the proteolysis catalyzed by the RgpB gingipain.

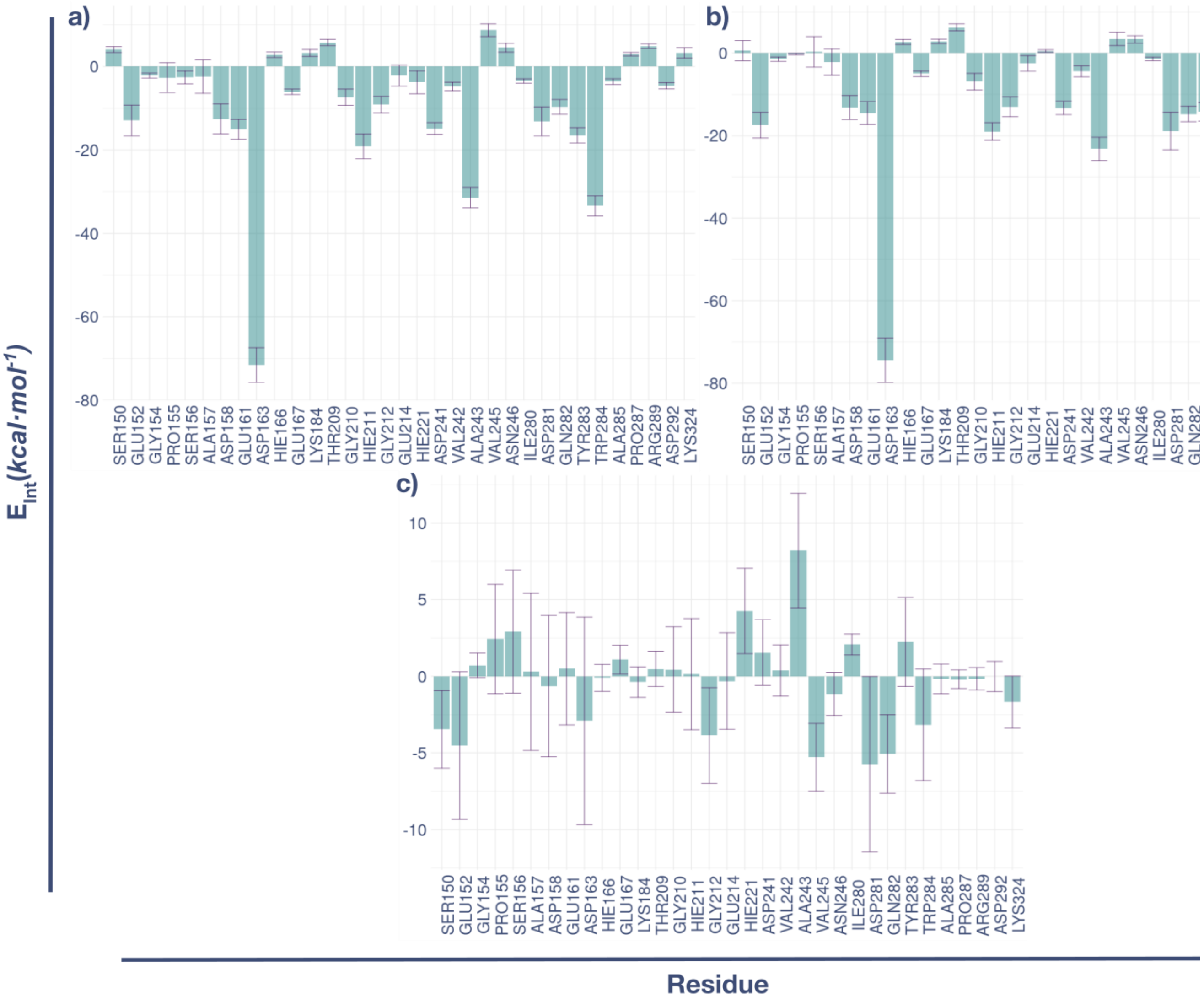


**Figure S4.** Main average interaction energies (electrostatic plus Lennard-Jones) between peptide and the protein residues along 1 ns of PM3/MM MD simulations at a) the **R1** reactant state, b) the **TS**$^{\mathbf{R1\text{-}I4}}$ transition state and c) the difference between **TS**$^{\mathbf{R1\text{-}I4}}$ and **R1**.

**Table S3.** Quantum region CHelpG charges on key atoms calculated on the localized PBE+D3(BJ)/MM stationary points.

**Acylation Stage**

| **Atom** | **R1** | $\mathbf{TS^{R1\rightarrow I4}}$ | **Δ** |
|---|---|---|---|
| Sγ:Cys244 | -0.7544 | -1.0501 | -0.2957 |
| Hγ:Cys244 | 0.3048 | 0.3880 | 0.0832 |
| C1:Pep | 0.3304 | 0.4686 | 0.1382 |
| N1:Pep | -0.3621 | -0.5654 | -0.2033 |
| Nδ:H211 | -0.4038 | -0.4091 | -0.0053 |

**Deacylation Stage**

| **Atom** | **I5** | $\mathbf{TS^{I5\rightarrow P1}}$ | **Δ** |
|---|---|---|---|
| Sγ:Cys244 | -0.4158 | -0.4563 | -0.0405 |
| C1:Pep | 0.8786 | 0.8569 | -0.0217 |
| O:Wat | -0.7043 | -0.5974 | 0.1069 |
| H:Wat | 0.4475 | 0.3736 | -0.0739 |
| Nδ:H211 | -0.6966 | -0.7014 | -0.0048 |

**Quantum region TS Cartesian Coordinates localized by PBE+D3(BJ)/MM calculations.**

**TS$^{R1 \rightarrow I4}$**

81

| | | | |
|---|---|---|---|
| C | 42.84886700 | 62.49327500 | 37.46338200 |
| H | 43.39067300 | 61.95185100 | 36.67355400 |
| H | 43.43118000 | 63.40327400 | 37.67508500 |
| C | 42.82209400 | 61.67642600 | 38.72051000 |
| N | 42.30584300 | 62.15851600 | 39.92202400 |
| C | 42.52604100 | 61.20384400 | 40.82072900 |
| H | 42.26839300 | 61.25477600 | 41.87685600 |
| N | 43.15583700 | 60.13484900 | 40.25834600 |
| H | 43.42341900 | 59.26963600 | 40.72128200 |
| C | 43.35429300 | 60.41960200 | 38.92009900 |
| H | 43.84452500 | 59.71848600 | 38.24987100 |
| C | 38.62810500 | 68.55434100 | 36.77846700 |
| O | 37.87948700 | 69.48394100 | 37.09991500 |
| N | 39.97469300 | 68.58706400 | 36.93065300 |
| H | 40.52237900 | 67.78981100 | 36.61266500 |
| C | 40.67658200 | 69.66669700 | 37.59104600 |
| H | 39.95294900 | 70.15262200 | 38.26311700 |
| C | 41.81537400 | 69.03051800 | 38.40665500 |
| H | 42.32135700 | 68.29294000 | 37.75888500 |
| H | 42.57260400 | 69.77980600 | 38.66718800 |
| S | 41.23192900 | 68.23054100 | 39.96085300 |
| H | 42.87656200 | 67.31791200 | 40.28822300 |
| C | 41.23696100 | 70.67396000 | 36.54233300 |
| O | 42.07384100 | 70.29867200 | 35.71917000 |
| N | 40.79248900 | 71.96291700 | 36.67674100 |
| H | 40.01864900 | 72.09621100 | 37.32338500 |
| C | 41.31709400 | 73.19463300 | 36.00741800 |
| H | 41.10662500 | 74.00666800 | 36.72830100 |
| C | 42.85175400 | 73.10239100 | 35.80390500 |
| H | 43.03091500 | 72.34118600 | 35.02921400 |
| C | 43.46794500 | 74.43292600 | 35.35112500 |
| H | 43.19809000 | 74.72267000 | 34.32669800 |
| H | 43.18939500 | 75.25684700 | 36.02901100 |
| H | 44.56546100 | 74.35705000 | 35.36736500 |
| C | 43.56381000 | 72.65566900 | 37.08267700 |
| H | 43.17817300 | 71.71493500 | 37.48916200 |
| H | 44.63047200 | 72.49482700 | 36.87089600 |
| H | 43.47642100 | 73.42575500 | 37.86644900 |
| C | 41.67834500 | 64.72674300 | 43.58285100 |
| O | 41.75292300 | 63.50376900 | 43.45279200 |
| N | 41.53410500 | 65.55511000 | 42.48479900 |
| H | 41.26432800 | 66.52784300 | 42.64282700 |
| C | 41.18857500 | 64.97729400 | 41.18769300 |
| H | 41.51752600 | 63.91493000 | 41.21051500 |
| C | 39.66674800 | 65.01737600 | 40.97048000 |
| H | 39.23999500 | 64.61300300 | 41.90738200 |
| H | 39.34420200 | 66.06997900 | 40.88702900 |
| C | 39.14528300 | 64.21829700 | 39.77028000 |

| | | | |
|---|---|---|---|
| H | 39.36637700 | 64.76475300 | 38.84186700 |
| H | 39.67027800 | 63.25000000 | 39.70062800 |
| C | 37.62984900 | 63.99519100 | 39.86238800 |
| H | 37.11696100 | 64.91058000 | 40.20041800 |
| H | 37.18534400 | 63.74222700 | 38.89035800 |
| N | 37.26911700 | 62.92326500 | 40.80947700 |
| H | 37.51163800 | 63.07243200 | 41.78705400 |
| C | 36.74677200 | 61.73165500 | 40.50175100 |
| N | 36.68144400 | 61.30880000 | 39.22889300 |
| H | 37.35249500 | 61.62098500 | 38.53648800 |
| H | 36.12391400 | 60.49551900 | 38.98565700 |
| N | 36.23700700 | 60.94378700 | 41.46798200 |
| H | 35.97716300 | 61.30715200 | 42.37803900 |
| H | 36.12828700 | 59.94366500 | 41.33281300 |
| C | 42.03223900 | 65.52632100 | 40.05356800 |
| O | 41.96712300 | 65.16449200 | 38.90790500 |
| N | 43.29281500 | 66.27160400 | 40.51777800 |
| H | 43.34805800 | 66.20509000 | 41.54255100 |
| C | 44.56319700 | 65.86035100 | 39.85930000 |
| H | 44.27631500 | 65.46568200 | 38.86999700 |
| C | 45.45678500 | 67.11180500 | 39.67207000 |
| H | 45.70071800 | 67.51878300 | 40.66616200 |
| C | 46.75859500 | 66.84653700 | 38.92152500 |
| H | 47.25352400 | 67.80661200 | 38.71796900 |
| H | 47.45836300 | 66.22943700 | 39.50311800 |
| H | 46.58677700 | 66.34374100 | 37.95761400 |
| O | 44.69879000 | 68.15567600 | 39.05704100 |
| H | 44.60578100 | 67.96577800 | 38.10355900 |
| H | 44.99156800 | 65.13044900 | 40.39197400 |
| H | 41.72823900 | 65.17438900 | 44.47566900 |
| H | 40.83564700 | 73.42924200 | 35.16292600 |
| H | 38.32058700 | 67.69326600 | 36.37352200 |
| H | 41.94635500 | 62.76049000 | 37.12564300 |

**TS$^{I5\rightarrow P1}$**

89

| | | | |
|---|---|---|---|
| C | 46.54474500 | 68.01315700 | 38.19335300 |
| H | 46.96405800 | 68.99563300 | 38.47284100 |
| H | 46.99524200 | 67.74778900 | 37.22513200 |
| C | 46.91384400 | 67.01856500 | 39.24060900 |
| N | 46.28405800 | 67.03331100 | 40.47135000 |
| C | 46.87159400 | 66.09310500 | 41.19157700 |
| H | 46.64707100 | 65.84539800 | 42.22464800 |
| N | 47.83627000 | 65.45848200 | 40.47123500 |
| H | 48.44481900 | 64.72037500 | 40.80945100 |
| C | 47.88349100 | 66.03774900 | 39.21792900 |
| H | 48.59385000 | 65.71628100 | 38.46148800 |
| C | 40.46848400 | 73.51246200 | 38.04350700 |
| O | 39.47055700 | 74.22597000 | 37.89963000 |
| N | 41.60724900 | 73.92330600 | 38.66658900 |
| H | 42.43511300 | 73.32866800 | 38.62399200 |
| C | 41.77915300 | 75.26714900 | 39.18044700 |
| H | 40.80410600 | 75.61957400 | 39.54085300 |
| C | 42.81333400 | 75.23908000 | 40.30731200 |
| H | 43.00122100 | 76.25505000 | 40.68154400 |
| H | 43.75469500 | 74.85783800 | 39.88384100 |
| S | 42.35847500 | 74.19752700 | 41.75138200 |
| C | 42.36647600 | 76.17875100 | 38.06296600 |
| O | 43.36085500 | 75.77806900 | 37.46160100 |
| N | 41.78559900 | 77.39636900 | 37.89055800 |
| H | 40.88628000 | 77.53905100 | 38.34412900 |
| C | 42.21440100 | 78.44384100 | 36.91731600 |
| H | 41.77571200 | 79.36965000 | 37.32294100 |
| C | 43.75868600 | 78.60067100 | 36.92930000 |
| H | 44.18564300 | 77.63490900 | 36.62442700 |
| C | 44.28650600 | 79.66403600 | 35.96092800 |
| H | 44.08748500 | 79.42382100 | 34.90666500 |
| H | 43.88538200 | 80.66542100 | 36.17365400 |
| H | 45.38051900 | 79.73217300 | 36.05769400 |
| C | 44.22997900 | 78.90990200 | 38.35398000 |
| H | 44.06295800 | 78.06855600 | 39.04151600 |
| H | 45.30366400 | 79.14611200 | 38.36961600 |
| H | 43.69148000 | 79.78238900 | 38.75807000 |
| N | 44.45384500 | 69.88635700 | 46.23421900 |
| H | 45.31770400 | 69.44934800 | 45.93437500 |
| C | 43.97817100 | 70.97104800 | 45.43415000 |
| H | 43.05036700 | 71.35970300 | 45.87871000 |
| C | 44.98081600 | 72.17882100 | 45.30689400 |
| H | 44.58692300 | 72.91444900 | 44.58457700 |
| H | 45.91516400 | 71.79154000 | 44.86600400 |
| C | 45.25359200 | 72.84132400 | 46.62913000 |
| C | 46.45906200 | 72.64454900 | 47.31657600 |
| H | 47.25318900 | 72.04965200 | 46.85482100 |
| C | 46.66981800 | 73.18799700 | 48.58872500 |
| H | 47.61902500 | 73.01150600 | 49.10714200 |

| | | | |
|---|---|---|---|
| C | 45.65717200 | 73.94017200 | 49.19983500 |
| O | 45.75696700 | 74.45013400 | 50.46742900 |
| H | 46.64233700 | 74.25818000 | 50.82745600 |
| C | 44.46763700 | 74.19824200 | 48.50288700 |
| H | 43.71167400 | 74.83472600 | 48.96727400 |
| C | 44.27733100 | 73.65106800 | 47.23673000 |
| H | 43.34501700 | 73.87064200 | 46.70592600 |
| C | 43.67613300 | 70.48342100 | 44.00281800 |
| O | 44.15200500 | 69.45550300 | 43.53205200 |
| N | 42.86439700 | 71.29586000 | 43.23064700 |
| H | 42.54358500 | 72.18400200 | 43.61050800 |
| C | 42.80979900 | 71.10505300 | 41.80004600 |
| H | 43.45739300 | 70.22703500 | 41.61225300 |
| C | 41.38440900 | 70.85050600 | 41.28742500 |
| H | 41.41989900 | 70.71466400 | 40.19623100 |
| H | 40.78777500 | 71.75785100 | 41.47142600 |
| C | 40.71979200 | 69.64728800 | 41.96047600 |
| H | 39.62673500 | 69.69947800 | 41.83501400 |
| H | 40.92570400 | 69.70688800 | 43.04348200 |
| C | 41.22304000 | 68.31120700 | 41.41190800 |
| H | 42.32693000 | 68.26624100 | 41.42799900 |
| H | 40.87463900 | 68.17176800 | 40.37637100 |
| N | 40.68691500 | 67.20876800 | 42.22807400 |
| H | 40.64283600 | 67.39392800 | 43.23052800 |
| C | 40.55804300 | 65.93430600 | 41.85111700 |
| N | 40.93327800 | 65.53635500 | 40.62377400 |
| H | 41.69474700 | 66.01211300 | 40.14905900 |
| H | 40.72858800 | 64.59340600 | 40.30701100 |
| N | 39.99998200 | 65.02500300 | 42.67611700 |
| H | 39.54558000 | 65.30714100 | 43.53884500 |
| H | 40.26789900 | 64.04787300 | 42.61956700 |
| C | 43.45994000 | 72.21996200 | 40.99224600 |
| O | 43.69287800 | 72.21653000 | 39.80900300 |
| O | 44.69499200 | 72.82518000 | 41.94962600 |
| H | 43.94968400 | 73.75199200 | 42.13384300 |
| H | 45.39711100 | 73.07078700 | 41.31140300 |
| H | 44.21084500 | 69.69794800 | 47.18577200 |
| H | 41.82039900 | 78.31799300 | 36.00686300 |
| H | 40.52354400 | 72.56899200 | 37.71665400 |
| H | 45.55298300 | 68.09643400 | 38.09602900 |

# ANNEX B. ARTICLE 2

# Computational Study of the Inhibition of RgpB Gingipain, a Promising Target for the Treatment of Alzheimer's Disease

Santiago Movilla[1], Sergio Martí[1], Maite Roca[1,*] and Vicent Moliner[1,*]

[1]BioComp Group, Institute of Advanced Materials (INAM), Universitat Jaume I, 12071, Castellón, Spain.

---

**ABSTRACT:** Alzheimer's disease represents one of the most ambitious challenges for biomedical sciences due to the growing number of cases in the world-wide elderly population and the lack of efficient treatments. One of the recent attempts to develop a treatment points to the cysteine protease RgpB as a promising drug target. In this attempt, several small molecule covalent inhibitors of this enzyme have been proposed. Here, we report a computational study at the atomic level of the inhibition mechanism of the most promising reported compounds. Molecular dynamic simulations were performed on six of them, and their binding energies in the active site of the protein were computed. Contact maps and interaction energies were decomposed by residues to disclose those key interactions with the enzyme. Finally, quantum mechanics / molecular mechanics (QM/MM) MD simulations were performed to evaluate the reaction mechanism by which these drug candidates lead to covalently bound complexes, inhibiting the RgpB protease. The results provide a guide for future redesign of prospective and efficient inhibitors for the treatment of Alzheimer's disease.

---

## 1. INTRODUCTION

Gingipains are a group of enzymes secreted by the pathogenic bacterium *Porphyromonas gingivalis.*[1] Although their activity is mainly related to oropharyngeal problems, there are reports that also associate them with other health disorders.[2–5] Some of these disorders are of particular medical interest, such as heart conditions and Alzheimer's disease.[2,6–11] The increasing number of neurodegenerative diseases worldwide and the lack of an appropriate medical treatment explains why this is the greatest current challenges for biomedical sciences.

Gingipains are a small family of cysteine proteases enzymes which catalyze the cleavage of peptidic bonds in a several number of proteic substrates.[12] In this family, we can find two types of gingipains, Lys-gingipains (Kgp) and Arg-gingipains (Rgp), classified according to the residue that they recognize at the P1 position, lysine and arginine, respectively, to perform the cleavage.[12] Without extra marked preferences for other amino acids at positions beyond P1, gingipains catalyze the proteolysis of a large number of proteins and peptides.[13,14] This makes them potentially harmful to the host integrity. Indeed, a recent study strongly linked one of the Rgp, RgpB, to the progression of Alzheimer's disease.[6] This study also showed that RgpB inhibition resulted in a strong arrest of disease progression in mice. With this study, RgpB was positioned as an important pharmacological target to understand its function and selectively inhibit it.

The proteolysis reaction mechanism of RgpB was recently studied by computational methods and reported in detail by our group.[15] In this study, we highlight several critical features to understand the catalytic mechanism of this cysteine protease and proceed to its inhibition. In general, and as is common in cysteine proteases, the mechanism is divided into an acylation step and a deacylation step.[16–26] According to our results,[15] the deprotonation of the nucleophilic cysteine of RgpB in the acylation stage is performed by the substrate itself in response to the steric impossibility of the catalytic histidine to perform this process. Similar mechanisms for activation have been reported in enzymes whose active site distribution coincides with that of RgpB,[18] where the substrate is positioned between the Cys/His catalytic dyad. After the deprotonation of the catalytic Cys, the sulfur atom of Cys attacks on the carbonyl group of the substrate to obtain the acylenzyme. The deacylation stage is carried out in a single step where a water molecule attacks the carbonyl group of the substrate and one proton of the water is transferred to the sulfur atom of Cys residue assisted by the catalytic His.

The interest in the inhibition of gingipains has been growing since their potential implications in the treatment of neurodegenerative diseases has been demonstrated.[27] In fact, some potential drug candidates, which have been patented (International Patent Application PCT/US2015/054050 and PCT/US2016/061197), are in advanced stages of medical trials. Interestingly, these compounds were reported to be irreversible inhibitors after the corresponding kinetic tests had been carried out.[6] Given the pharmacological potential of these molecules, the main goal of the present study is to investigate how these irreversible inhibitors can stop the catalytic activity of RgpB as protease.

Structurally, these inhibitors have the side chain pattern of an arginine but without the nitrogen of the P1-P1' peptide bond. The absence of this nitrogen makes impossible to carry out the hydrolysis reaction and breaking the peptide bond, thus inhibiting the enzyme. Interestingly, they also lack the reactive warheads commonly used for covalent inhibition of cysteine proteases.[20,28–30] This fact limits the reactive possibilities of these compounds against the action of RgpB gingipain. In our previously reported study,[15] it was proposed that, in the absence of a peptidic nitrogen, the catalytic cysteine activation was carried out by the proton transfer from the sulfur atom of Cys to the oxygen atom of the carbonyl group of the substrate.[15] Based on these results, we proposed that this mechanism could be exploited for the design of potential irreversible drugs.

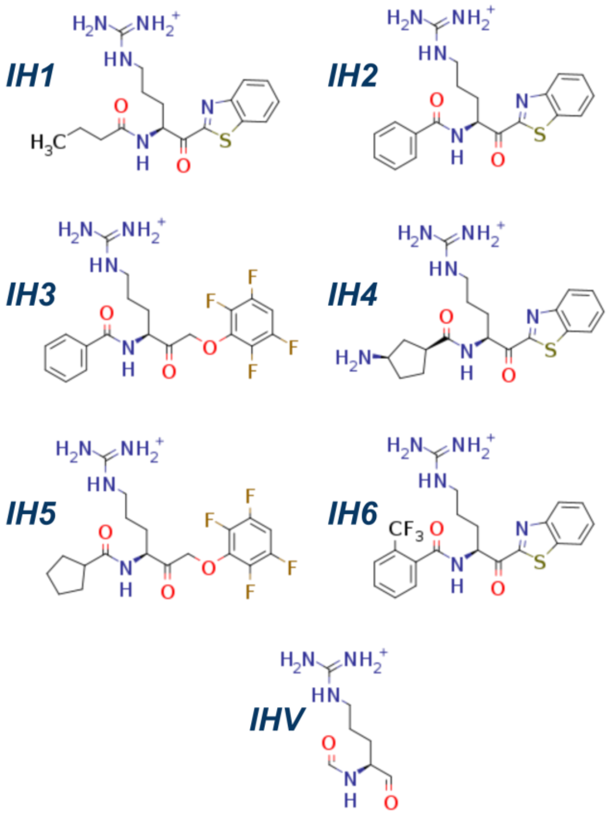
This background and in view of the limited reactive nature of the already irreversible patented inhibitors, a detailed understanding of their inhibition mechanism is necessary to exploit their potential in the future. Namely, in the case of covalent inhibitors, the inhibitory potency derives from the synergy between the non-covalent interactions with the enzyme and the kinetics/thermodynamics of the reactive process.[29] Therefore, an atomic-level study of the inhibition mechanism must incorporate both the study of the inhibitor-enzyme interactions and the study of the covalent binding reaction mechanism. Here, we present a comprehensive computational study that provides insights into important atomic details in the reaction mechanism of inhibition, binding affinity and the interaction between the enzyme and a representative group of inhibitors. Classical molecular dynamics (MD) simulations, alchemical transformations,[31] molecular mechanics Poisson-Boltzmann surface area (MM/PBSA)[32] calculations supplemented with interaction entropies, and MD simulations with quantum mechanics molecular mechanics (QM/MM) potentials were carried out to get a detailed description of the inhibitor-enzyme binding step and the chemical steps of the covalent inhibitor-enzyme bond formation. The obtained results throw light on the design of better inhibitors for the Alzheimer's disease treatment.

## 2. METHODS

**System Set Up.** The initial coordinates of the system were obtained from the previously equilibrated model[15] of RgpB in complex with a short peptide. This model was reported in the reactivity study of RgpB performed by our group. The peptide was replaced by six active inhibitors, (Scheme 1) reported in International Patent Application PCT/US2016/061197, that covered the largest extent of the chemical space. The protonation state of all titratable residues was verified using the PROPKA3 server[33] at reference pH 7.5,[6,13,14] no atypical protonation states were observed. Each of the systems was solvated with a cubic box of TIP3P water[34] molecules with a minimum distance of 15 Å between any solute atom and the edge of the box. A total number of 18 $Na^+$ ions were added until the systems were neutralized.

**Classical Molecular Dynamics (MD).** Protein and water molecules were described using the AMBERff14SB[35] and TIP3P[34] force field parameters, respectively. Monovalent ions were treated using the parameters proposed by Joung-Cheatham.[36] For inhibitors, restrained electrostatic potential (RESP) charges[37] were obtained from HF/6-31G* calculations on optimized geometries with B3LYP functional[38,39] with the 6-31G* basis set, according to previously reported protocols.[40] The remaining force field parameters for the inhibitors were assigned from the small-molecules generalized Amber force field (GAFF).[40] The topologies of each system were obtained using the tLEaP package from Ambertools20.[41]

**SCHEME 1. Chemical structures of the studied irreversible inhibitors of RgpB (IH1-IH6) and sketch of the virtual inhibitor (IHV) used as common point for the alchemical transformations.**



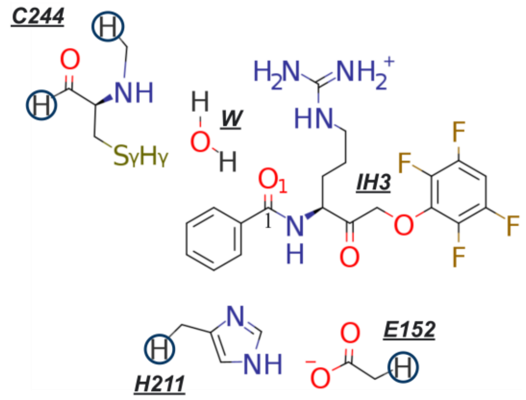
All MD classical simulations were carried out in several steps: (i) solvent molecules, ions and hydrogens were minimized using 2500 steps with the conjugate gradient algorithm; (ii) 200 ps of MD simulations of the solvent molecules and mono-valent ions were carried out, with the positions of the backbone atoms restrained; (iii) Two energy minimizations, one with the protein backbone restrained and another fully unrestrained were done for each system; (iv) The whole systems were heated gradually in three consecutive NPT simulations from 100 K to 310 K, with a constant pressure of 1 bar; and finally, (v) 250 ns of NPT MD simulations were performed at 310 K and 1 bar. All parameters of the classical simulations were replicated from the previously reported study.[15] A 10 Å cutoff was set up for short-range non-bonded interactions while a Particle Mesh Ewald (PME)[42,43] model was used for the long-range interactions. Langevin dynamics thermostat[44,45] was applied with a 3 $ps^{-1}$ collision frequency. For all equilibration simulations, the SHAKE algorithm[46,47] was used to constrain light atoms, and the velocity Verlet algorithm[48] was used to propagate the system.

**MM/PBSA and Interaction Entropies.** To estimate the binding free energy ($\Delta G_{bind}$) of the inhibitors, calculations of MM/PBSA[32] supplemented with corrections to the solute entropy using the methodology proposed by Duan et. al.[49] were performed. For these calculations, frames were taken every 500 ps from each classical MD simulation. MM/PBSA calculations were performed with an implicit salt concentration of 150 mM (to reproduce experimental conditions)[6,13] and using MMPBSA.py as implemented in Amber20.[41] In order to obtain the binding free energy ($\Delta G_{bind}$), we applied the interaction entropy approach[49] to compute the entropic contribution term ($-T\Delta S$) at 310 K.

**Alchemical Calculations.** In order to obtain accurate relative values of the binding free energy ($\Delta\Delta G_{bind}$), we performed alchemical transformations[31] from the inhibitors to a virtual ligand (IHV) connecting them all (Scheme 1). For this purpose, the 3-step Amber Thermodynamic Integration protocol ("decharge-LJ-recharge" protocol)[50] was used. Each transformation was carried out in ten windows equally distributed throughout the $\lambda$ range (0-1). In each window, 500 ps of CPU equilibration (pmemd.MPI) and 5 ns of sampling with GPU algorithms (pmemd.cuda) were performed.[51,52] Only the atoms appearing/disappearing during the transformation were included in the soft-core region. All the simulations were performed at NVT ensemble at 310 K, starting from the volume equilibrated structures of the classical MD trajectories. For data analysis, the first 5% of the simulation time (250 ps) of each window was discarded.

**Potential of Mean Force (PMF).** The six studied inhibitors present the same reactive carbonyl warhead (C1:IH3 – O1:IH3, Scheme 2). Then, the inhibitor that showed the most favorable binding energy was selected for studying the reaction mechanism of the formation of the inhibitor-enzyme covalent complex by generating the free energy surfaces (FES) in terms of PMFs. For the PMF calculations, 2D potential energy surfaces (PES) were first computed through sequential minimizations along selected collective variables (CV) that best describe the reactions. A conjugate gradient algorithm was employed for the minimizations using a gradient tolerance of 0.1 $\text{kcal mol}^{-1}$ as a convergence criterion. Later, the corresponding FES were generated at 310 K taking as starting points the structures obtained along the PES. Each window had a relaxation time of 5 ps and a sampling time of 25 ps using a time step of 0.5 fs in NVT ensemble. Temperature control was performed using Langevin dynamics with a 3 $\text{ps}^{-1}$ collision frequency.[44,45] The umbrella sampling method[53] was used to restrain the reaction coordinates. The force constant used for each window was 580 $\text{kcal·mol}^{-1}\text{·Å}^{-2}$, and the window width was 0.05 Å for those collective variables corresponding to the antisymmetric combination of two distances and 0.1 Å for those that are distances. The number and the width of the windows selected ensure a correct overlapping of windows. The umbrella integration method, implemented in the QM3 suite,[54] was used to reweight the biased sampling dynamics and to generate the PMFs along selected coordinates. The quantum mechanical (QM) region used for the generation of the PMFs is shown in Scheme 2. The method used to describe the QM region was PM6.[55] Following the same protocol of our previous study on this system,[15] a structure close to the transition state region was selected and the transition state structure was located at PM6/MM level and verified by tracing down the intrinsic reaction coordinate path (IRC). The obtained minima were optimized. Single point energy calculations were carried out on the stationary point structures at the PBE-D3(BJ)[56–60]/MM level with the 6-311+G** basis set, to correct the PM6/MM potential energy. The thermal contributions calculated by the statistical methods at the PM6/MM level were thus preserved. In order to verify the results, all the critical points were reoptimized at the PBE-D3(BJ)/MM level using QM3 suite.[54]

**SCHEME 2. Schematic representation of the region treated quantum mechanically to explore the inhibitory mechanism of binding between IH3 and RgpB. The link atoms are shown as hydrogens inside circles. The water molecule (W) was included in the QM region when exploring the mechanism with participation of a water molecule.**



## 3. RESULTS AND DISCUSSION

**Binding affinities ($\Delta G_{bind}$) and inhibitor-enzyme interaction profiles.** In order to study the non-covalent interaction profiles between inhibitors and RgpB enzyme, 250 ns of classical MD simulations were run per inhibitor in the inhibitor-enzyme non-covalent reactant complex. None of the systems, in the presence of the corresponding inhibitor, exhibited significant changes in the protein structure (all backbone RMSDs < 2 Å on average, Figure S1).

**TABLE 1. Binging free energies for the studied inhibitors obtained over the classical MD simulations using MM/PBSA and interaction entropies ($\Delta G_{bind-MM/PBSA}$) and binging free energies relative to the virtual inhibitor (IHV) computed by means of alchemical transformations ($\Delta G_{bind-FEP}$). All values are in kcal·mol$^{-1}$.**

| Inhibitor | $\Delta H_{bind-MM/PBSA}$ | $-T\Delta S_{bind-MM/PBSA}$ | $\Delta G_{bind-MM/PBSA}$ | $\Delta\Delta G_{bind-FEP}$ |
|---|---|---|---|---|
| **IH1** | -57.3 | 37.0 | -20.3 | 0.20 |
| **IH2** | -54.2 | 50.6 | -3.6 | 5.42 |
| **IH3** | -69.4 | 27.3 | -42.1 | 0.55 |
| **IH4** | -67.2 | 28.2 | -39.0 | 1.99 |
| **IH5** | -63.3 | 27.8 | -35.5 | 2.55 |
| **IH6** | -60.9 | 33.1 | -27.8 | 3.31 |

Starting from the equilibrated structures of the classical MD simulations, we initially proceeded to estimate the binding free energies for each of the compounds. For this purpose, we computed the binding affinity energies by MM/PBSA[32] calculations supplemented with a correction term to the entropic contributions based on the interaction entropies proposed by Duan et al.[49] (Table 1). Qualitatively speaking, compounds IH3 and IH4 proved to produce the most stable inhibitor-enzyme complexes. It should also be noted that IH3 is the one that shows the most favorable binding energy in both entropic and enthalpic terms. On the other hand, inhibitor IH2 presented a lower binding enthalpy and a higher -TΔS term. We emphasize that the values computed by means of MM/PBSA and interaction entropies have no quantitative meaning and are purely used to analyze qualitatively the enthalpic and entropic contributions of the binding processes.[61] Thus, keeping in mind the inherent uncertainty associated to the MM/PBSA method, alchemical transformations[31] were employed to compute the differences between the binding free energies of every inhibitor and the virtual inhibitor IHV ($\Delta\Delta G_{bind-TI}$) in order to obtain more precise and quantitatively meaningful inhibitor-enzyme affinities.

The results of the alchemical transformations show a similar trend to the one obtained from MM/PBSA calculations, with the only exception of IH1, previously ranked as the second weakest, is repositioned in first place with practically the same $\Delta\Delta G_{bind-TI}$ as IH3. Same as the MM/PBSA calculations,[32] IH4 ranks after IH3 as one of the most potent candidates. Likewise, IH2 is the inhibitor with the lowest affinity to the enzyme with a difference of 5.2 kcal·mol$^{-1}$ with respect to IH1. Henceforth, IH3 can be considered as the reference inhibitor given the agreement between the binding free energies estimated by both methods.

A contact frequency map allowed us to analyze the differences in the interaction patterns/profiles of each inhibitor, which can be complemented with the analysis of the averaged inhibitor-enzyme interaction energies decomposed by residue (Figure 1). Figure 1a shows the relative contact frequencies with respect to the inhibitor with the highest affinity, IH3. Thus, positive values represent a higher contact frequency than in IH3 while negative values represent a lower contact frequency than in IH3. In general, most of the inhibitors present a similar interaction profile with the enzyme, which is in agreement with the differences observed between the binding free energies ($\Delta\Delta G_{bind-TI}$) computed by alchemical transformations. However, it can be observed from the contact map that three inhibitors (IH2, IH4 and IH6) show a significantly lower contact frequency with residues Ser213 and Glu214 (see Figure 1a). A visual inspection revealed that for the case of inhibitors IH3 and IH5 this interaction corresponds to a hydrogen bond between the HN:Glu214 and the F1 atom of the inhibitor. In the case of the interaction with Ser213 residue does not reach a hydrogen bond due to the angle of the atoms involved (angle around 59±16 degrees) (Figure 2). Thus, for those inhibitors that instead of the tetrafluorophenoxy substituent have the benzothiazole ring the interactions with Ser213 and Glu214 are not present. However, IH1 manages to interact with HN:Glu214 via the nitrogen atom of the bicyclo ring. Although no preferences in selectivity of the groups around arginine moiety have been reported in RgpB that affect the catalytic capacity, it has shown a higher affinity for hydrophobic substituents, mainly aromatic.[13] However, these results suggest that hydrogen bonding groups are necessary to stabilize these aromatic substituents within the active site and can be optimized to achieve better affinity to the enzyme.

Another interaction worth highlighting is the one observed between compound IH4 and residues Asp158 and Asp281 (see Figure 1a and Figure 2). Namely, IH4 is the only one of the compounds studied that presents a hydrogen bond donor at the nitrogen substituent group of the arginine. Consequently, HN3:IH4 manages to interacts with Oδ:Asp158 (Figure 2). As a result, this inhibitor is repositioned in such a way that it manages to form an extra hydrogen bond interaction between the HN1:IH4 and O:Asp281 atoms. Although this interaction would be possible in all inhibitors, it was only observed in IH4, suggesting that it responds to the conformation adopted due to the interaction with residue Asp158. Other differences in the contact map, such as

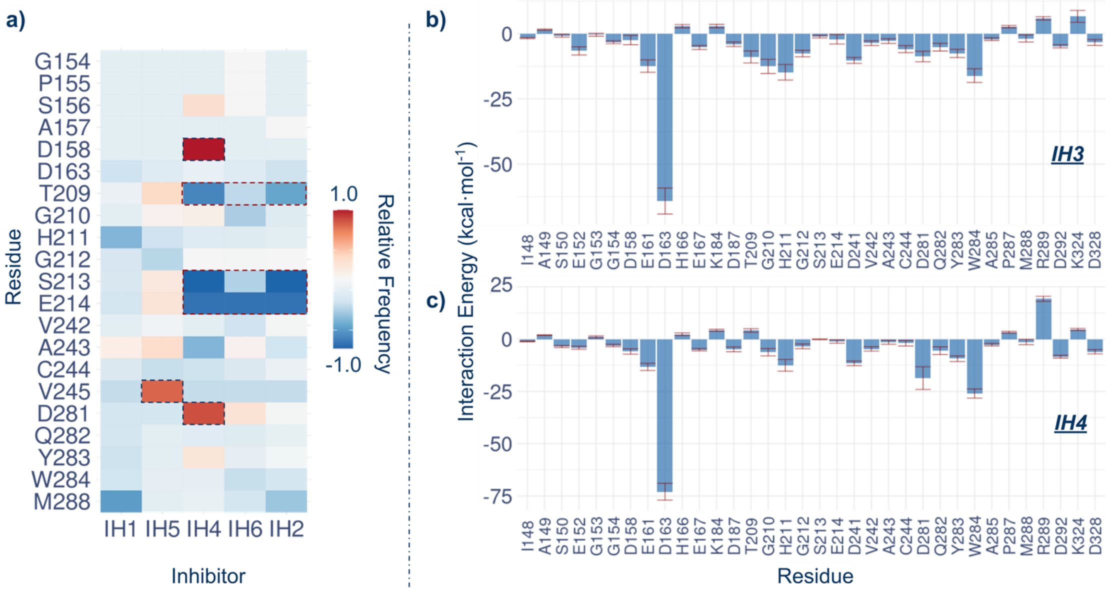


**FIGURE 1. a) Relative frequency of contacts between residues of RgpB and the inhibitors. IH3 was used as reference and a contact was counted if the distance between atoms was < 4.0 Å. The cells corresponding to the interactions that differ the most from the IH3 inhibitor are highlighted with dashed-line borders. Panels b) and c) show averaged interaction energies (electrostatic plus Lennard–Jones) between residues of RgpB and IH3 and IH4, respectively, over 1 ns MD simulations at PM3/MM level.**

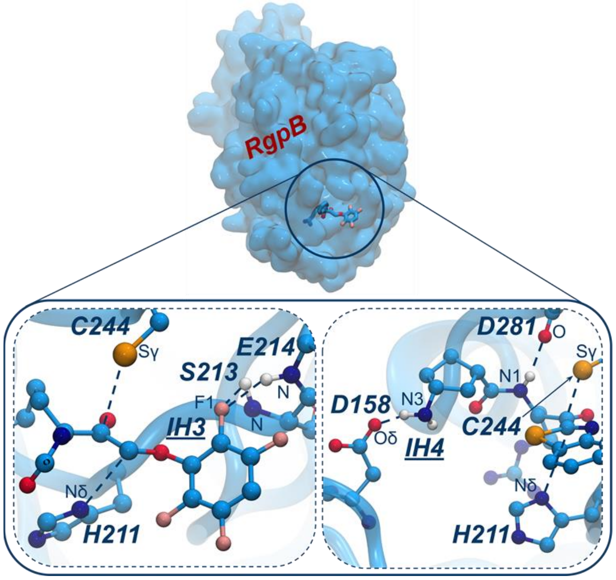


**FIGURE 2. Top panel: surface representation of the whole protein and location of the binding pocket. IH3 inhibitor is represented as sticks. Bottom panels: Schematic representation of the key interactions between IH3 (left) or IH4 (right) and the binding pocket of RgpB.**

those observed in residues Thr209 or Val245, are less specific and are the consequence of a particular physical proximity during the simulations.

Finally, in order to energetically characterize the favorable observed contacts between the inhibitors and the RgpB residues, the interaction energies were calculated (Figures 1b, 1c, S3, S4, S5 and S6). We can observe the determinant role played by residues Asp163, Trp284 and His211 in arginine binding, a conclusion that was predicted from structural analysis and our previous report on the proteolysis reaction catalyzed by RgpB.[15] This result is not surprising given the high selectivity of RgpB to arginine residues. On the other hand, it can be observed that the interaction with residues Ser213 and Glu214 is always favorable. In the same way, the interaction between IH4 and residues Asp158 and Asp281 shows a lower energy, being remarkable the one with Asp281. In contrast, residues Lys184, Arg289 and Lys324 show an unfavorable interaction energy with all compounds. Although this may be a starting point for future optimizations, it derives from the spatial proximity (without contact) between the charges of these residues and the positive charge of the guanidinium group of the arginine moiety of the inhibitors.

To sum up, our results suggest that the optimization should be focused on compounds presenting hydrogen bond acceptor groups capable to interact with HN:E214 and HN:Ser213. Moreover, hydrogen bond donor groups interacting with Asp158 and Asp281 seems to enhance the affinity with the receptor. On the other hand, the guanidinium group present in all inhibitors and in the natural substrate should be preserved as has been shown to be essential for recognition by RgpB enzyme.

**Covalent binding chemical step.** In order to study the reaction mechanism by which these inhibitors covalently bind to the enzyme, QM/MM MD simulations were performed to generate the full free energy landscape. Given the similarity between the inhibitors, the reactivity study was carried out using only the inhibitor IH3, which showed the most favorable binding energy.

Initially, and analogously to the proteolysis reaction catalyzed by the enzyme in the presence of wild-type substrate, Cys244 is protonated and requires activation to attack the carbonyl group in the reactant state (**R** in Figures 3 a) and b), and Figure 4 b)). We also considered the possibility that the mechanism proceeded starting from the deprotonated Sγ:C244 form (mechanism $R_{S(-)} \rightarrow P_{S(-)}$ in Figure 3 c)). However, several attempts (with and without restrains) to obtain stable reactive structures were unsuccessful. In all simulated cases, the negative charge on C244 resulted in deformations of the system to chemically unviable structures.

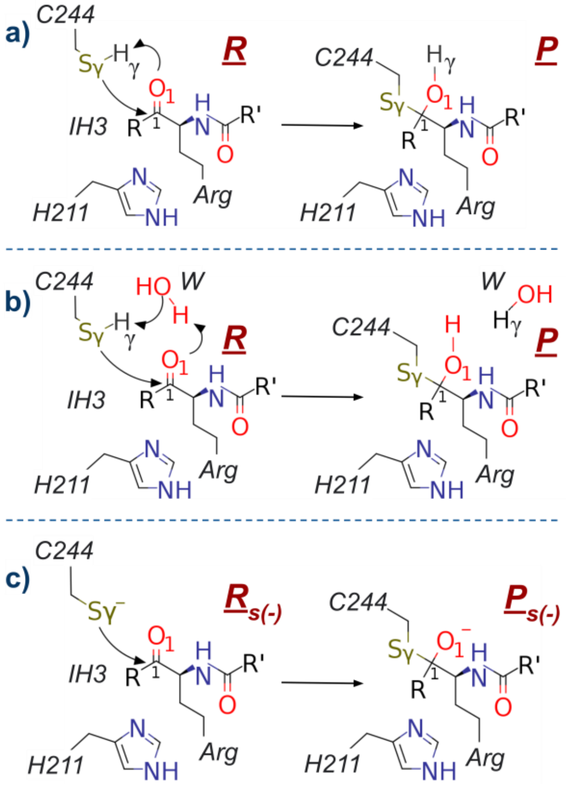


**FIGURE 3. Schematic representation of the considered mechanisms of covalent binding between IH3 and RgpB. Sketch of the reaction mechanism in which the Hγ:Cys244 transfer to the O1 atom of the inhibitor occurs a) directly or b) mediated by a water molecule. c) Representation of the reaction mechanism starting from the deprotonated Sγ:C244 form. As shown in Scheme1 for IH3, R and R' correspond to phenyl and 2,3,5,6-tetrafluorophenoxy substituents, respectively.**

The family of inhibitors reported does not possess a clear reactive moiety, such as a Michael acceptor or an epoxide group. Instead, they only possess the carbonyl oxygen atom O1:IH3 as a possible acceptor/activator for the cysteine residue. This possibility was previously hypothesized[15] in light of the small difference between the free energy barriers shown by the wild-type being activated by either the peptide nitrogen (more favorable mechanism by ~1.1 kcal·mol$^{-1}$) or carbonyl oxygen of the substrate. Other possible bases were evaluated by calculating the population of hydrogen bonds between Hγ:Cys244 and other possible proton acceptors. However, the interaction with O1:IH3 was shown to populate more than 50% of the time, while other possible bases, such as the Asp281, interacted less than 1% with it, making them poor base candidates. With the carbonyl group as the only plausible activator of Cys244, the mechanisms are limited to whether the attack of Sγ:Cys244 on C1:IH3 and the transfer of Hγ:Cys244 from Sγ:Cys244 to O1:IH3 occur directly or mediated by a water molecule (Figure 3 a) and b)).

As revealed by the computed FESs (Figure S7), both mechanisms, the one in which the Hγ:Cys244 is directly transferred to the O1:IH3 atom (mechanism a in Fig. 3), or the one mediated by a water molecule (mechanism b in Fig. 3), proceed through a single concerted step. In the case of the direct mechanism, that was explored using the distance Sγ:C244-C1:IH3 and the difference between distances [Sγ:C244-Hγ:C244] - [O1:IH3-Hγ:C244] as collective variables to describe the process, the activation free energy associated to the transition state (**TS**, see Figure 4 a)) is 22.8 kcal·mol$^{-1}$ with respect to the reactants. A value that is very close to that previously computed for the natural substrate, 23.5 kcal·mol$^{-1}$.[15] However, in the case of the transition state mediated by a water molecule (**TS**$_W$), that was explored using the distance Sγ:C244-C1:IH3 and the difference between distances [Sγ:C244-Hγ:C244] - [O:W-Hγ:C244] as collective variables to describe the process, the corresponding activation energy is 14.2 kcal·mol$^{-1}$. This result suggests this mechanism as the most viable for the covalent binding reaction with a significantly lower free energy barrier compared to that of the wild-type substrate. The obtained product (**P**) is located at -12.1 kcal·mol$^{-1}$ with respect to the reactants, thus giving an irreversible exergonic reaction. In order to verify the mechanism obtained at PM6/MM level, **R**, **TS**$_W$ and **P** were fully optimized at a higher level of theory, PBE+D3(BJ)/MM. The high-level optimized structures, presented in Figure 4, match the reaction pathway predicted at lower level, verifying the mechanism deduced from the PM6/MM FES.

As mentioned above, the mechanism starting from the deprotonate C244 (mechanism c in Figure 3) was not explored because the active site was deformed along the classical MD simulations and no properly reactive structures were obtained.

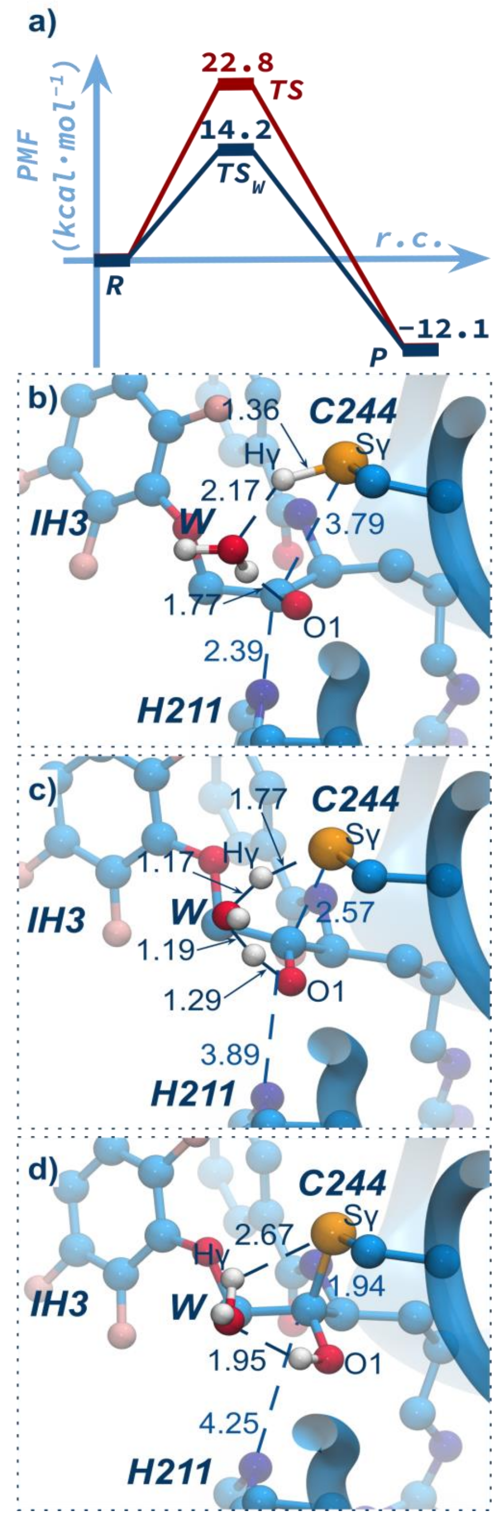


**FIGURE 4. a) Free energy profiles computed for the inhibitory covalent binding mechanisms of RgpB by IH3 at the PBE+D3(BJ):PM6/MM. PBE+D3(BJ)/MM optimized structures of the critical points b) R, c) $TS_W$ and d) P along the most likely covalent binding mechanism of IH3 inhibiting RgpB catalytic activity. Selected distances are in Å. FESs are shown in Figure S7.**

In light of these results, it is remarkable to highlight the ability of RgpB to react without the presence of highly reactive groups. Common reactive groups such as Michael acceptors or epoxides deal with selectivity issues due to their reactivity with other undesired targets. The unusual carbonyl warhead enables to exploit the design of covalent inhibitors, with high potency, without having to face problems of selectivity. These features qualify RgpB as a pharmacological target that promises effective treatments by the use of these kind of inhibitors without side effects for the treatment of Alzheimer's disease.

## 4. CONCLUSIONS

Alzheimer's disease is one of the most studied medical challenges today. Gingipains proteases, including RgpB, have positioned themselves as potential drug targets for the development of treatments for the disease. Herein, we unveil atomic-level details of the non-covalent binding processes for a set of RgpB gingipain inhibitors, potential candidates for the treatment of Alzheimer's disease. Six compounds, presenting the side chain pattern of an arginine but without the nitrogen of the P1-P1' peptide bond, were chosen in order to cover the widest breadth of the chemical space previously proposed and patented,[6] some of them are in advanced stages of clinical testing. Initially, non-covalent inhibitor-enzyme complexes were simulated by classical molecular dynamics. Over the MD sampling, binding free energies ($\Delta\Delta G_{bind\text{-}TI}$) were computed with alchemical transformation calculations.

Based on a relative contact map analysis between the inhibitors and the enzyme, we concluded that IH3, IH1 and IH5 present similar pattern of interactions, showing only minor differences between the contacts established with the protein residues along the sampled trajectories. Meanwhile, IH2, IH4 and IH6, that some of them rendered lower binding energies, do not interact with key residues such as Ser213 and Glu214. However, IH4 exhibits two hydrogen bond interactions that no other shows with residues Asp158 and Asp281. These interactions as well as those with Ser213 and Glu214 were found to have a favorable interaction energy whenever present. These analyses of the non-covalent complexes suggest that IH1 and IH3 would be the most promising candidates for further refinements, according to their high binding affinities to the active site of RgpB.

The reaction mechanism of the covalent bond formation between the inhibitors and the enzyme was computed by QM/MM MD simulations on IH3, which showed the best binding profile in the non-covalent complex. The most plausible reaction mechanism proceeds through a single concerted step in which activation/deprotonation of Cys244 is carried out by O1:IH3 atom mediated by a water molecule (W). At the same time, the Sγ:Cys244 atom attacks the C1:IH3 of the inhibitor to give way to a stable product (-12.1 kcal·mol$^{-1}$). The reaction proceeds through an activation barrier (14.2 kcal·mol$^{-1}$) significantly lower than that reported previously, computed in our laboratory for the wild-type substrate (23.5 kcal·mol$^{-1}$).[15] Thus, it is remarkable the ability of RgpB to react without the presence of highly reactive groups, enabling to exploit the re-design of covalent inhibitors, recognized for their potency, without predicting problems of selectivity. These results qualify RgpB as a pharmacological target that promises effective treatments by the use of these kind of inhibitors without side effects. In particular, the inhibitors should contain hydrogen bond donor and acceptor groups to be able to interact with Glu214, Ser213, Asp158 and Asp281, and conserving the reported guanidinium group and warhead to provide potency and selectivity. The interactions reported here and the mechanistic details represent a key starting point for future re-design of prospective and efficient inhibitors for the treatment of Alzheimer's disease.

## ACKNOWLEDGEMENTS

We would like to thank the Spanish Ministerio de Ciencia e Innovación (grant PGC2021-23332OB-C21), the Generalitat Valenciana (PROMETEO, grant ref. CIPROM/2021/079) and Universitat Jaume I (grants UJI-2020-03 and UJI-2019-43). Santiago Movilla thanks the Generalitat Valenciana for a Grisolia Ph.D. grant (GRISOLIAP/2019/064). Authors acknowledge computational resources from the Servei d'Informàtica of Universitat Jaume I.

## ASSOCIATED CONTENT

### SUPPORTING INFORMATION

Root-mean-square deviation (RMSD) analyses for classical MD simulations; root-mean-square fluctuation (RMSF) analyses for classical MD simulations; averaged interaction energies between the inhibitors and some protein residues for IH1, IH2, IH5, IH6 and IH7; Free energy surfaces for the covalent binding inhibition mechanisms of the RgpB gingipain with IH3; Force Field parameters for IH3 and IHV; Cartesian coordinates for R, TS$_W$ and P structures localized at PBE+D3(BJ)/MM level for the covalent binding mechanism with IH3 inhibitor.

## AUTHOR INFORMATION

### Corresponding Author

**Maite Roca** – BioComp Group, Institute of Advanced Materials (INAM), Universidad Jaume I, 12071 Castellón, Spain; orcid.org/0000-0003-0937-4722; Email: mroca@uji.es

**Vicent Moliner** – BioComp Group, Institute of Advanced Materials (INAM), Universidad Jaume I, 12071 Castellón, Spain; orcid.org/0000-0002-3665-3391; Email: moliner@uji.es

### Authors

**Santiago Movilla** – BioComp Group, Institute of Advanced Materials (INAM), Universidad Jaume I, 12071 Castellón, Spain.

**Sergio Martí** – BioComp Group, Institute of Advanced Materials (INAM), Universidad Jaume I, 12071 Castellón, Spain; orcid.org/0000-0002-1087-7143

### Notes

The authors declare no competing financial interest.

## REFERENCES

(1) Darveau, R. P.; Hajishengallis, G.; Curtis, M. A. Porphyromonas Gingivalis as a Potential Community Activist for Disease. *J. Dent. Res.* **2012**, *91* (9), 816–820. https://doi.org/10.1177/0022034512453589.

(2) Mahendra, J.; Mahendra, L.; Kurian, V. M.; Jaishankar, K.; Mythilli, R. Prevalence of Periodontal Pathogens in Coronary Atherosclerotic Plaque of Patients Undergoing Coronary Artery Bypass Graft Surgery. *J. Maxillofac. Oral Surg.* **2009**, *8* (2), 108–113. https://doi.org/10.1007/s12663-009-0028-5.

(3) Katz, J.; Chegini, N.; Shiverick, K. T.; Lamont, R. J. Localization of P. Gingivalis in Preterm Delivery Placenta. *J. Dent. Res.* **2009**, *88* (6), 575–578. https://doi.org/10.1177/0022034509338032.

(4) Ishikawa, M.; Yoshida, K.; Okamura, H.; Ochiai, K.; Takamura, H.; Fujiwara, N.; Ozaki, K. Oral Porphyromonas Gingivalis Translocates to the Liver and Regulates Hepatic Glycogen Synthesis through the Akt/GSK-3β Signaling Pathway. *Biochim. Biophys. Acta - Mol. Basis Dis.* **2013**, *1832* (12), 2035–2043. https://doi.org/10.1016/j.bbadis.2013.07.012.

(5) Mougeot, J.-L. C.; Stevens, C. B.; Paster, B. J.; Brennan, M. T.; Lockhart, P. B.; Mougeot, F. K. B. Porphyromonas Gingivalis Is the Most Abundant Species Detected in Coronary and Femoral Arteries. *J. Oral Microbiol.* **2017**, *9* (1), 1281562. https://doi.org/10.1080/20002297.2017.1281562.

(6) Dominy, S. S.; Lynch, C.; Ermini, F.; Benedyk, M.; Marczyk, A.; Konradi, A.; Nguyen, M.; Haditsch, U.; Raha, D.; Griffin, C.; Holsinger, L. J.; Arastu-Kapur, S.; Kaba, S.; Lee, A.; Ryder, M. I.; Potempa, B.; Mydel, P.; Hellvard, A.; Adamowicz, K.; Hasturk, H.; Walker, G. D.; Reynolds, E. C.; Faull, R. L. M.; Curtis, M. A.; Dragunow, M.; Potempa, J. Porphyromonas Gingivalis in Alzheimer's Disease Brains: Evidence for Disease Causation and Treatment with Small-Molecule Inhibitors. *Sci. Adv.* **2019**, *5* (1), eaau3333. https://doi.org/10.1126/sciadv.aau3333.

(7) Kaye, E. K.; Valencia, A.; Baba, N.; Spiro, A.; Dietrich, T.; Garcia, R. I. Tooth Loss and Periodontal Disease Predict Poor Cognitive Function in Older Men. *J. Am. Geriatr. Soc.* **2010**, *58* (4), 713–718. https://doi.org/10.1111/j.1532-5415.2010.02788.x.

(8) Gatz, M.; Mortimer, J. A.; Fratiglioni, L.; Johansson, B.; Berg, S.; Reynolds, C. A.; Pedersen, N. L. Potentially Mod-

ifiable Risk Factors for Dementia in Identical Twins. *Alzheimer's Dement.* **2006**, *2* (2), 110–117. https://doi.org/10.1016/j.jalz.2006.01.002.

(9) Stein, P. S.; Desrosiers, M.; Donegan, S. J.; Yepes, J. F.; Kryscio, R. J. Tooth Loss, Dementia and Neuropathology in the Nun Study. *J. Am. Dent. Assoc.* **2007**, *138* (10), 1314–1322. https://doi.org/10.14219/jada.archive.2007.0046.

(10) Kamer, A. R.; Pirraglia, E.; Tsui, W.; Rusinek, H.; Vallabhajosula, S.; Mosconi, L.; Yi, L.; McHugh, P.; Craig, R. G.; Svetcov, S.; Linker, R.; Shi, C.; Glodzik, L.; Williams, S.; Corby, P.; Saxena, D.; de Leon, M. J. Periodontal Disease Associates with Higher Brain Amyloid Load in Normal Elderly. *Neurobiol. Aging* **2015**, *36* (2), 627–633. https://doi.org/10.1016/j.neurobiolaging.2014.10.038.

(11) Noble, J. M.; Borrell, L. N.; Papapanou, P. N.; Elkind, M. S. V; Scarmeas, N.; Wright, C. B. Periodontitis Is Associated with Cognitive Impairment among Older Adults: Analysis of NHANES-III. *J. Neurol. Neurosurg. Psychiatry* **2009**, *80* (11), 1206–1211. https://doi.org/10.1136/jnnp.2009.174029.

(12) Mayrand, D.; Holt, S. C. Biology of Asaccharolytic Black-Pigmented Bacteroides Species. *Microbiol. Rev.* **1988**, *52* (1), 134–152. https://doi.org/10.1128/mmbr.52.1.134-152.1988.

(13) Ally, N.; Whisstock, J. C.; Sleprawska-Lupa, M.; Potempa, J.; Le Bonniec, B. F.; Travis, J.; Pike, R. N. Characterization of the Specificity of Arginine-Specific Gingipains from Porphyromonas Gingivalis Reveals Active Site Differences between Different Forms of the Enzymes. *Biochemistry* **2003**, *42* (40), 11693–11700. https://doi.org/10.1021/bi0349726.

(14) Eichinger, A.; Beisel, H. G.; Jacob, U.; Huber, R.; Medrano, F. J.; Banbula, A.; Potempa, J.; Travis, J.; Bode, W. Crystal Structure of Gingipain R: An Arg-Specific Bacterial Cysteine Proteinase with a Caspase-like Fold. *EMBO J.* **1999**, *18*, 5453–5462. https://doi.org/10.1093/emboj/18.20.5453.

(15) Movilla, S.; Martí, S.; Roca, M.; Moliner, V. Unrevealing the Proteolytic Activity of RgpB Gingipain from Computational Simulations. *J. Chem. Inf. Model.* **2021**, *61* (9), 4582–4593. https://doi.org/10.1021/acs.jcim.1c00666.

(16) Wei, D.; Huang, X.; Liu, J.; Tang, M.; Zhan, C. G. Reaction Pathway and Free Energy Profile for Papain-Catalyzed Hydrolysis of N-Acetyl-Phe-Gly 4-Nitroanilide. *Biochemistry* **2013**, *52* (30), 5145–5154. https://doi.org/10.1021/bi400629r.

(17) Świderek, K.; Moliner, V. Revealing the Molecular Mechanisms of Proteolysis of SARS-CoV-2 Mproby QM/MM Computational Methods. *Chem. Sci.* **2020**, *11* (39), 10626–10630. https://doi.org/10.1039/d0sc02823a.

(18) Elsässer, B.; Zauner, F. B.; Messner, J.; Soh, W. T.; Dall, E.; Brandstetter, H. 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.* **2017**, *7* (9), 5585–5593. https://doi.org/10.1021/acscatal.7b01505.

(19) Ramos-Guzmán, C. A.; Zinovjev, K.; Tuñón, I. Modeling Caspase-1 Inhibition: Implications for Catalytic Mechanism and Drug Design. *Eur. J. Med. Chem.* **2019**, *169*, 159–167. https://doi.org/10.1016/j.ejmech.2019.02.064.

(20) Arafet, K.; Ferrer, S.; González, F. V.; Moliner, V. Quantum Mechanics/Molecular Mechanics Studies of the Mechanism of Cysteine Protease Inhibition by Peptidyl-2,3-Epoxyketones. *Phys. Chem. Chem. Phys.* **2017**, *19* (20), 12740–12748. https://doi.org/10.1039/C7CP01726J.

(21) Arafet, K.; Świderek, K.; Moliner, V. Computational Study of the Michaelis Complex Formation and the Effect on the Reaction Mechanism of Cruzain Cysteine Protease. *ACS Omega* **2018**, *3* (12), 18613–18622. https://doi.org/10.1021/acsomega.8b03010.

(22) Sulpizi, M.; Rothlisberger, U.; Carloni, P. Molecular Dynamics Studies of Caspase-3. *Biophys. J.* **2003**, *84* (4), 2207–2215. https://doi.org/10.1016/S0006-3495(03)75026-7.

(23) Miscione, G. Pietro; Calvaresi, M.; Bottoni, A. Computational Evidence for the Catalytic Mechanism of Caspase-7. A DFT Investigation. *J. Phys. Chem. B* **2010**, *114* (13), 4637–4645. https://doi.org/10.1021/jp908991z.

(24) Angelides, K. J.; Fink, A. L. Mechanism of Thiol Protease Catalysis: Detection and Stabilization of a Tetrahedral Intermediate in Papain Catalysis. *Biochemistry* **1979**, *18* (11), 2363–2369. https://doi.org/10.1021/bi00578a035.

(25) Ŝtrajbl, M.; Florián, J.; Warshel, A. Ab Initio Evaluation of the Free Energy Surfaces for the General Base/Acid Catalyzed Thiolysis of Formamide and the Hydrolysis of Methyl Thiolformate: A Reference Solution Reaction for Studies of Cysteine Proteases. *J. Phys. Chem. B* **2001**, *105* (19), 4471–4484. https://doi.org/10.1021/jp010279l.

(26) Arafet, K.; Ferrer, S.; Moliner, V. Computational Study of the Catalytic Mechanism of the Cruzain Cysteine Protease. *ACS Catal.* **2017**, *7* (2), 1207–1215. https://doi.org/10.1021/acscatal.6b03096.

(27) Parra-Torres, V.; Melgar-Rodríguez, S.; Muñoz-Manríquez, C.; Sanhueza, B.; Cafferata, E. A.; Paula-Lima, A. C.; Díaz-Zúñiga, J. Periodontal Bacteria in the Brain—Implication for Alzheimer's Disease: A Systematic Review. *Oral Dis.* **2021**. https://doi.org/10.1111/odi.14054.

(28) Martí, S.; Arafet, K.; Lodola, A.; Mulholland, A. J.; Świderek, K.; Moliner, V. Impact of Warhead Modulations on the Covalent Inhibition of SARS-CoV-2 M pro Explored by QM/MM Simulations. *ACS Catal.* **2022**, *12* (1), 698–708. https://doi.org/10.1021/acscatal.1c04661.

(29) Gehringer, M.; Laufer, S. A. Emerging and Re-Emerging Warheads for Targeted Covalent Inhibitors: Applications in Medicinal Chemistry and Chemical Biology. *J. Med. Chem.* **2019**, *62* (12), 5673–5724. https://doi.org/10.1021/acs.jmedchem.8b01153.

(30) Arafet, K.; Serrano-Aparicio, N.; Lodola, A.; Mulholland, A. J.; González, F. V.; Świderek, K.; Moliner, V. Mechanism of Inhibition of SARS-CoV-2 M pro by N3 Peptidyl Michael Acceptor Explained by QM/MM Simulations and Design of New Derivatives with Tunable Chemical

Reactivity. *Chem. Sci.* **2021**, *12* (4), 1433–1444. https://doi.org/10.1039/D0SC06195F.

(31) Jorgensen, W. L.; Ravimohan, C. Monte Carlo Simulation of Differences in Free Energies of Hydration. *J. Chem. Phys.* **1985**, *83* (6), 3050–3054. https://doi.org/10.1063/1.449208.

(32) Kollman, P. A.; Massova, I.; Reyes, C.; Kuhn, B.; Huo, S.; Chong, L.; Lee, M.; Lee, T.; Duan, Y.; Wang, W.; Donini, O.; Cieplak, P.; Srinivasan, J.; Case, D. A.; Cheatham, T. E. Calculating Structures and Free Energies of Complex Molecules: Combining Molecular Mechanics and Continuum Models. *Acc. Chem. Res.* **2000**, *33* (12), 889–897. https://doi.org/10.1021/ar000033j.

(33) Olsson, M. H. M.; Søndergaard, C. R.; Rostkowski, M.; Jensen, J. H. PROPKA3: Consistent Treatment of Internal and Surface Residues in Empirical p K a Predictions. *J. Chem. Theory Comput.* **2011**, *7* (2), 525–537. https://doi.org/10.1021/ct100578z.

(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.* **1983**, *79* (2), 926–935. https://doi.org/10.1063/1.445869.

(35) Maier, J. A.; Martinez, C.; Kasavajhala, K.; Wickstrom, L.; Hauser, K. E.; Simmerling, C. Ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from Ff99SB. *J. Chem. Theory Comput.* **2015**, *11* (8), 3696–3713. https://doi.org/10.1021/acs.jctc.5b00255.

(36) Joung, I. S.; Cheatham, T. E. Determination of Alkali and Halide Monovalent Ion Parameters for Use in Explicitly Solvated Biomolecular Simulations. *J. Phys. Chem. B* **2008**, *112* (30), 9020–9041. https://doi.org/10.1021/jp8001614.

(37) Woods, R.; Chappelle, R. Restrained Electrostatic Potential Atomic Partial Charges for Condensed-Phase Simulations of Carbohydrates. *J. Mol. Struct. THEOCHEM* **2000**, *527* (1–3), 149–156. https://doi.org/10.1016/S0166-1280(00)00487-5.

(38) Becke, A. D. Density-Functional Exchange-Energy Approximation with Correct Asymptotic Behavior. *Phys. Rev. A* **1988**, *38* (6), 3098–3100. https://doi.org/10.1103/PhysRevA.38.3098.

(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* **1988**, *37* (2), 785–789. https://doi.org/10.1103/PhysRevB.37.785.

(40) Wang, J.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A. Development and Testing of a General Amber Force Field. *J. Comput. Chem.* **2004**, *25* (9), 1157–1174. https://doi.org/10.1002/jcc.20035.

(41) Salomon-Ferrer, R.; Case, D. A.; Walker, R. C. An Overview of the Amber Biomolecular Simulation Package. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2013**, *3* (2), 198–210. https://doi.org/10.1002/wcms.1121.

(42) Essmann, U.; Perera, L.; Berkowitz, M. L.; Darden, T.; Lee, H.; Pedersen, L. G. A Smooth Particle Mesh Ewald Method. *J. Chem. Phys.* **1995**, *103* (19), 8577–8593. https://doi.org/10.1063/1.470117.

(43) Darden, T.; York, D.; Pedersen, L. Particle Mesh Ewald: An N ·log( N ) Method for Ewald Sums in Large Systems. *J. Chem. Phys.* **1993**, *98* (12), 10089–10092. https://doi.org/10.1063/1.464397.

(44) Uberuaga, B. P.; Anghel, M.; Voter, A. F. Synchronization of Trajectories in Canonical Molecular-Dynamics Simulations: Observation, Explanation, and Exploitation. *J. Chem. Phys.* **2004**, *120* (14), 6363–6374. https://doi.org/10.1063/1.1667473.

(45) Sindhikara, D. J.; Kim, S.; Voter, A. F.; Roitberg, A. E. Bad Seeds Sprout Perilous Dynamics: Stochastic Thermostat Induced Trajectory Synchronization in Biomolecules. *J. Chem. Theory Comput.* **2009**, *5* (6), 1624–1631. https://doi.org/10.1021/ct800573m.

(46) Ryckaert, J.-P.; Ciccotti, G.; Berendsen, H. J. Numerical Integration of the Cartesian Equations of Motion of a System with Constraints: Molecular Dynamics of n-Alkanes. *J. Comput. Phys.* **1977**, *23* (3), 327–341. https://doi.org/10.1016/0021-9991(77)90098-5.

(47) Kräutler, V.; van Gunsteren, W. F.; Hünenberger, P. H. A Fast SHAKE Algorithm to Solve Distance Constraint Equations for Small Molecules in Molecular Dynamics Simulations. *J. Comput. Chem.* **2001**, *22* (5), 501–508. https://doi.org/10.1002/1096-987X(20010415)22:5<501::AID-JCC1021>3.0.CO;2-V.

(48) Verlet, L. Computer "Experiments" on Classical Fluids. I. Thermodynamical Properties of Lennard-Jones Molecules. *Phys. Rev.* **1967**, *159* (1), 98–103. https://doi.org/10.1103/PhysRev.159.98.

(49) Duan, L.; Liu, X.; Zhang, J. Z. H. Interaction Entropy: A New Paradigm for Highly Efficient and Reliable Computation of Protein–Ligand Binding Free Energy. *J. Am. Chem. Soc.* **2016**, *138* (17), 5722–5728. https://doi.org/10.1021/jacs.6b02682.

(50) Loeffler, H. H.; Bosisio, S.; Duarte Ramos Matos, G.; Suh, D.; Roux, B.; Mobley, D. L.; Michel, J. Reproducibility of Free Energy Calculations across Different Molecular Simulation Software Packages. *J. Chem. Theory Comput.* **2018**, *14* (11), 5567–5582. https://doi.org/10.1021/acs.jctc.8b00544.

(51) Götz, A. W.; Williamson, M. J.; Xu, D.; Poole, D.; Le Grand, S.; Walker, R. C. Routine Microsecond Molecular Dynamics Simulations with AMBER on GPUs. 1. Generalized Born. *J. Chem. Theory Comput.* **2012**, *8* (5), 1542–1555. https://doi.org/10.1021/ct200909j.

(52) 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.* **2013**, *9* (9), 3878–3888. https://doi.org/10.1021/ct400314y.

(53) Torrie, G. M.; Valleau, J. P. Nonphysical Sampling Distributions in Monte Carlo Free-Energy Estimation: Umbrella Sampling. *J. Comput. Phys.* **1977**, *23* (2), 187–199. https://doi.org/10.1016/0021-9991(77)90121-8.

(54) Martí, S. QMCube (QM3): An All-Purpose Suite for Multiscale QM/MM Calculations. *J. Comput. Chem.* **2021**, *42* (6), 447–457. https://doi.org/10.1002/jcc.26465.

(55) Stewart, J. J. P. Optimization of Parameters for Semiempirical Methods V: Modification of NDDO Approximations and Application to 70 Elements. *J. Mol. Model.* **2007**, *13* (12), 1173–1213. https://doi.org/10.1007/s00894-007-0233-4.

(56) Perdew, J. P.; Burke, K.; Ernzerhof, M. Generalized Gradient Approximation Made Simple. *Phys. Rev. Lett.* **1996**, *77* (18), 3865–3868. https://doi.org/10.1103/PhysRevLett.77.3865.

(57) Ditchfield, R.; Hehre, W. J.; Pople, J. A. Self-Consistent Molecular-Orbital Methods. IX. An Extended Gaussian-Type Basis for Molecular-Orbital Studies of Organic Molecules. *J. Chem. Phys.* **1971**, *54* (2), 724–728. https://doi.org/10.1063/1.1674902.

(58) Hehre, W. J.; Ditchfield, R.; Pople, J. A. Self—Consistent Molecular Orbital Methods. XII. Further Extensions of Gaussian—Type Basis Sets for Use in Molecular Orbital Studies of Organic Molecules. *J. Chem. Phys.* **1972**, *56* (5), 2257–2261. https://doi.org/10.1063/1.1677527.

(59) Clark, T.; Chandrasekhar, J.; Spitznagel, G. W.; Schleyer, P. V. R. Efficient Diffuse Function-Augmented Basis Sets for Anion Calculations. III. The 3-21+G Basis Set for First-Row Elements, Li-F. *J. Comput. Chem.* **1983**, *4* (3), 294–301. https://doi.org/10.1002/jcc.540040303.

(60) Grimme, S.; Ehrlich, S.; Goerigk, L. Effect of the Damping Function in Dispersion Corrected Density Functional Theory. *J. Comput. Chem.* **2011**, *32* (7), 1456–1465. https://doi.org/10.1002/jcc.21759.

(61) Hou, T.; Wang, J.; Li, Y.; Wang, W. Assessing the Performance of the MM/PBSA and MM/GBSA Methods. 1. The Accuracy of Binding Free Energy Calculations Based on Molecular Dynamics Simulations. *J. Chem. Inf. Model.* **2011**, *51* (1), 69–82. https://doi.org/10.1021/ci100275a.

# Supporting Information

# Computational Study of the Inhibition of RgpB Gingipain, a Promising Target for the Treatment of Alzheimer's Disease

Santiago Movilla[1], Sergio Martí[1], Maite Roca[1,*] and Vicent Moliner[1,*]

[1]BioComp Group, Institute of Advanced Materials (INAM), Universitat Jaume I, 12071, Castellón, Spain

[*]Correspondence should be addressed to:
Maite Roca: mroca@uji.es
Vicent Moliner: moliner@uji.es

# Table of Contents

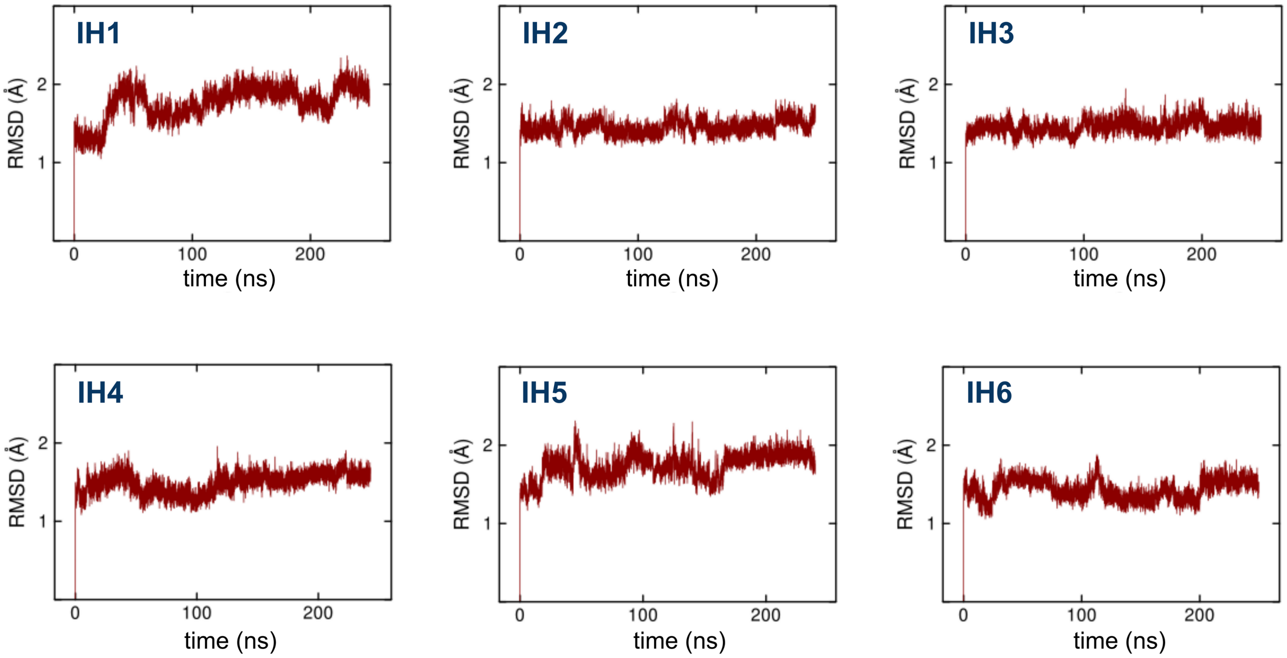


**Figure S1.** Time evolution of the root-mean-square deviation (RMSD) of protein backbone atoms (Cα, C, N, O) of the RgpB gingipain in complex with the six selected inhibitors (IH1 to IH6).

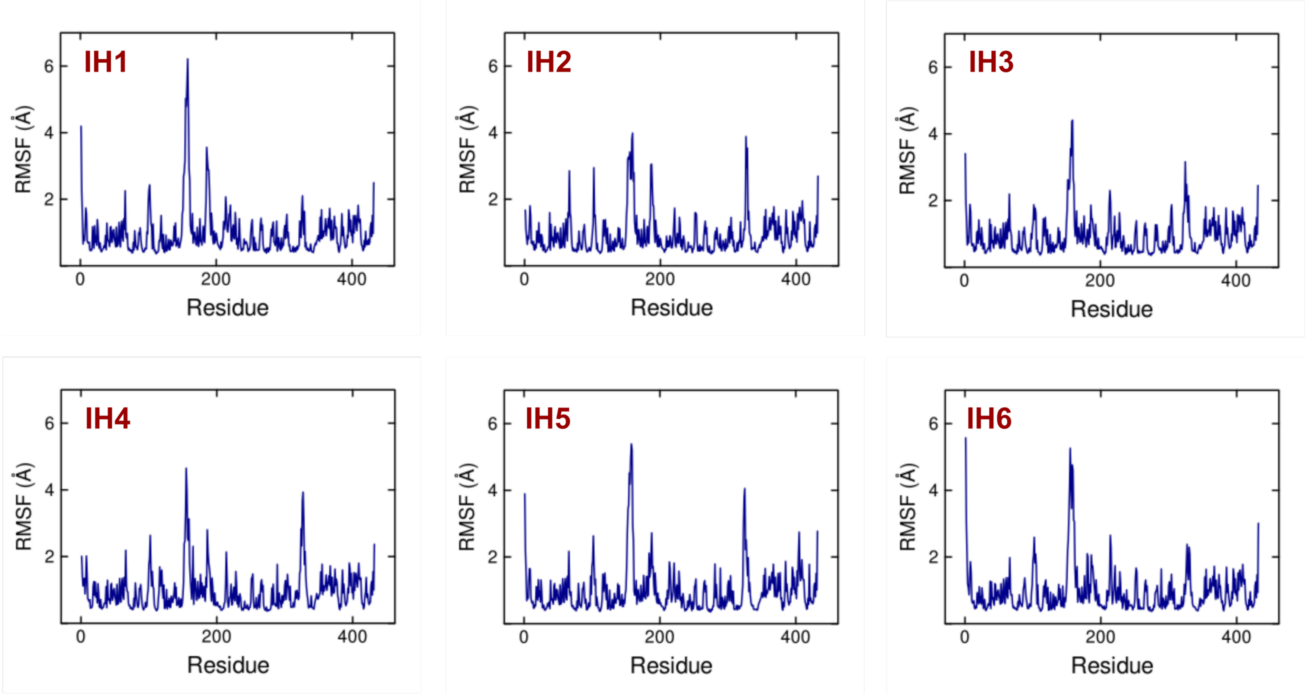


**Figure S2.** Root-mean-square fluctuation (RMSF) analysis of protein α carbon atoms of the RgpB gingipain in complex with the six selected inhibitors (IH1 to IH6).

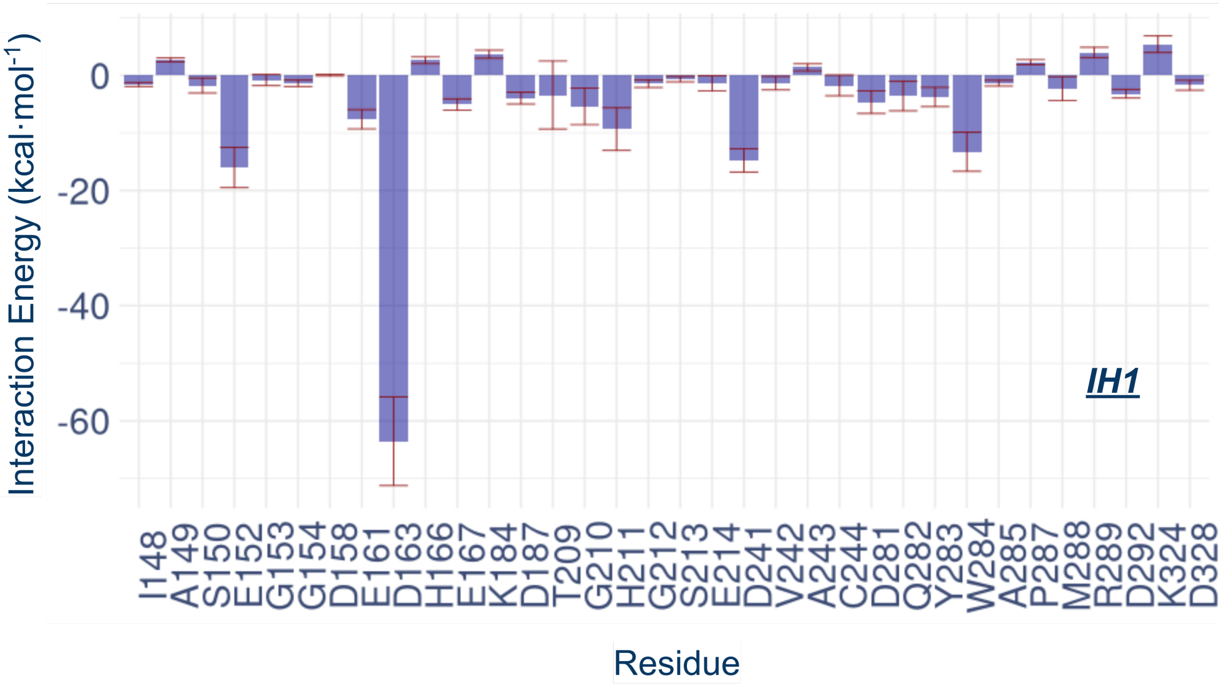


**Figure S3.** Averaged interaction energies (electrostatic plus Lennard−Jones) between some residues of RgpB and IH1.

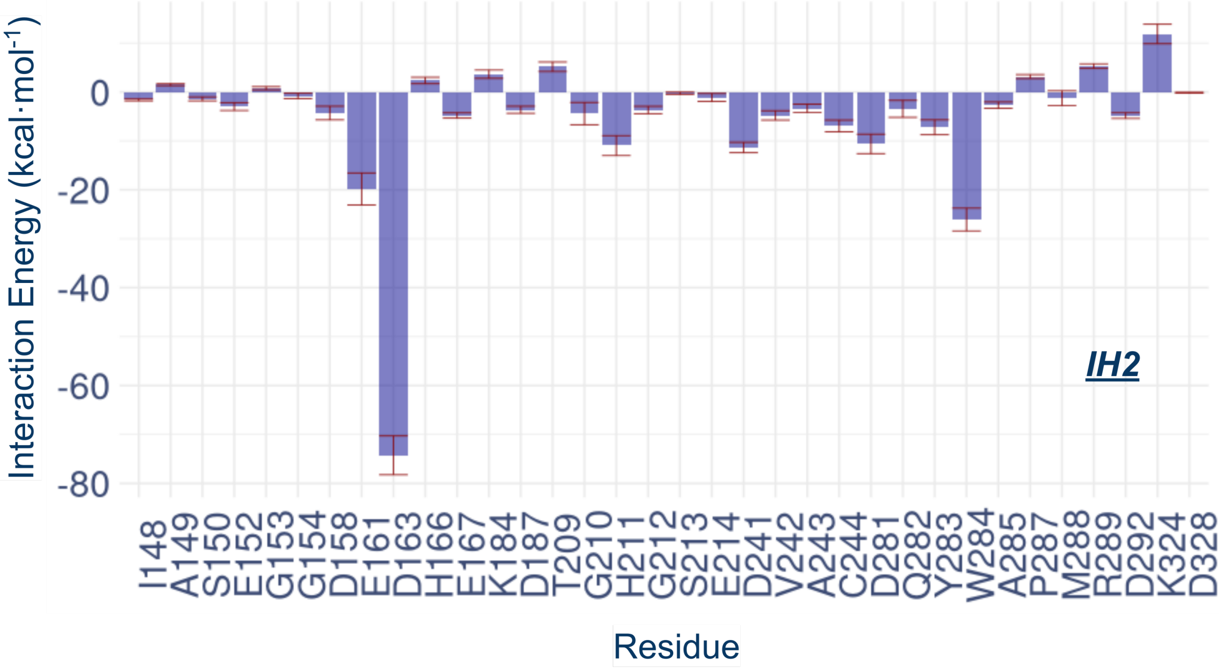


**Figure S4.** Averaged interaction energies (electrostatic plus Lennard−Jones) between some residues of RgpB and IH2.

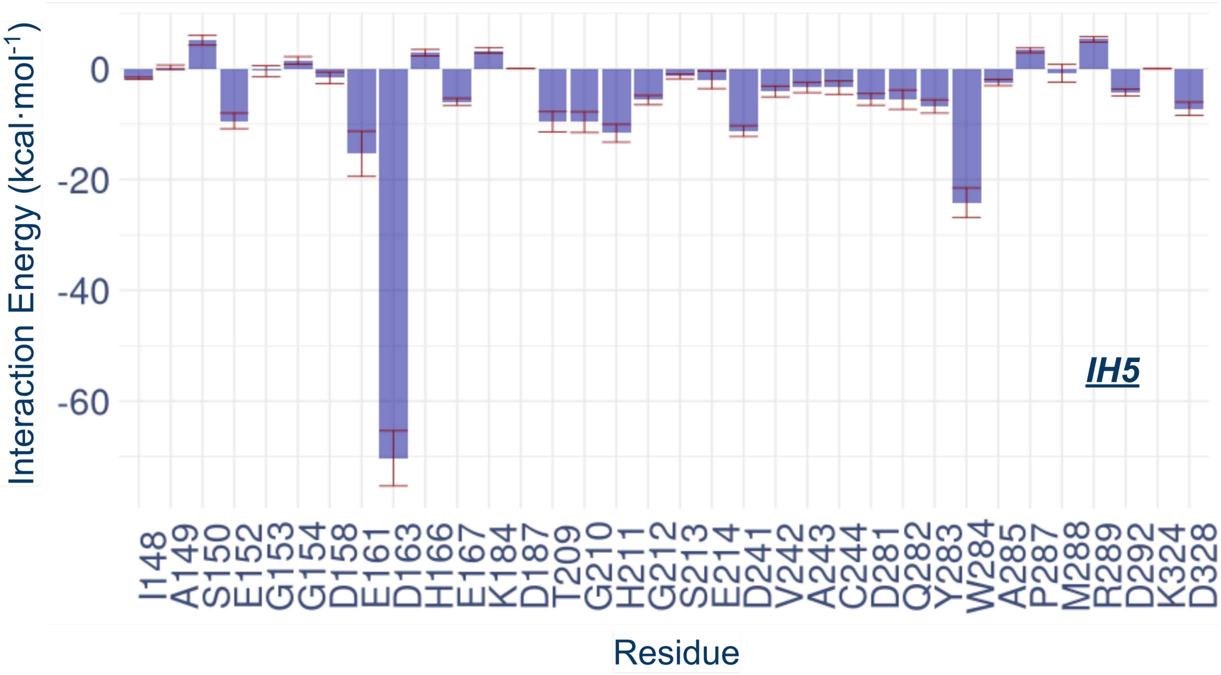


**Figure S5.** Averaged interaction energies (electrostatic plus Lennard−Jones) between some residues of RgpB and IH5.

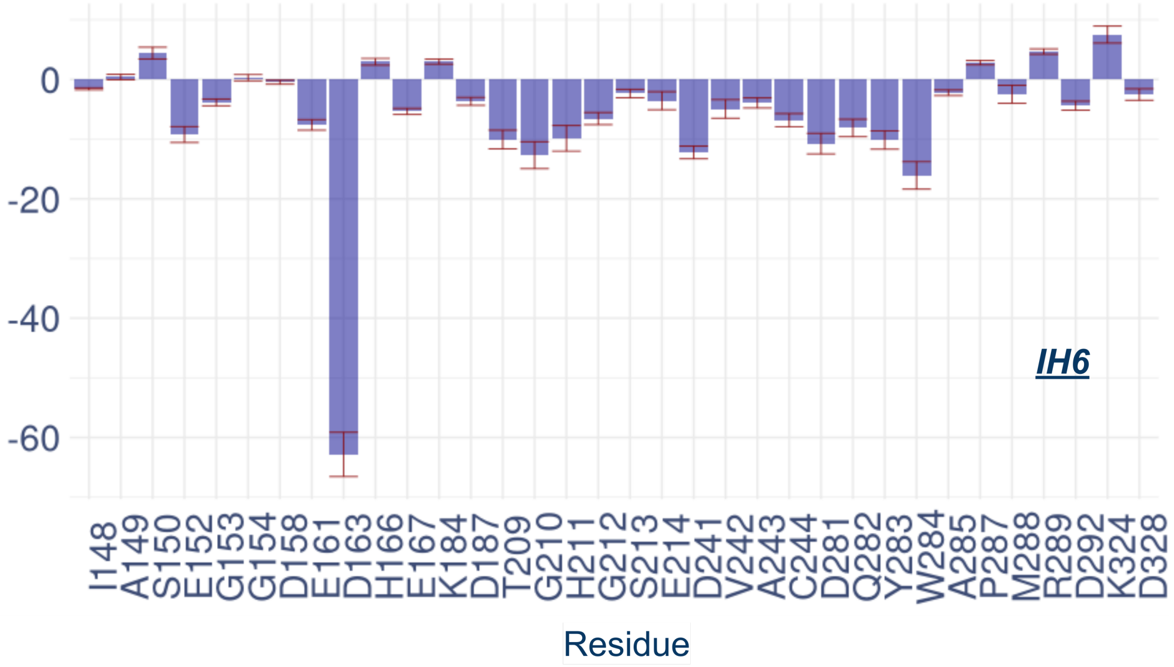


**Figure S6.** Averaged interaction energies (electrostatic plus Lennard−Jones) between some residues of RgpB and IH6.

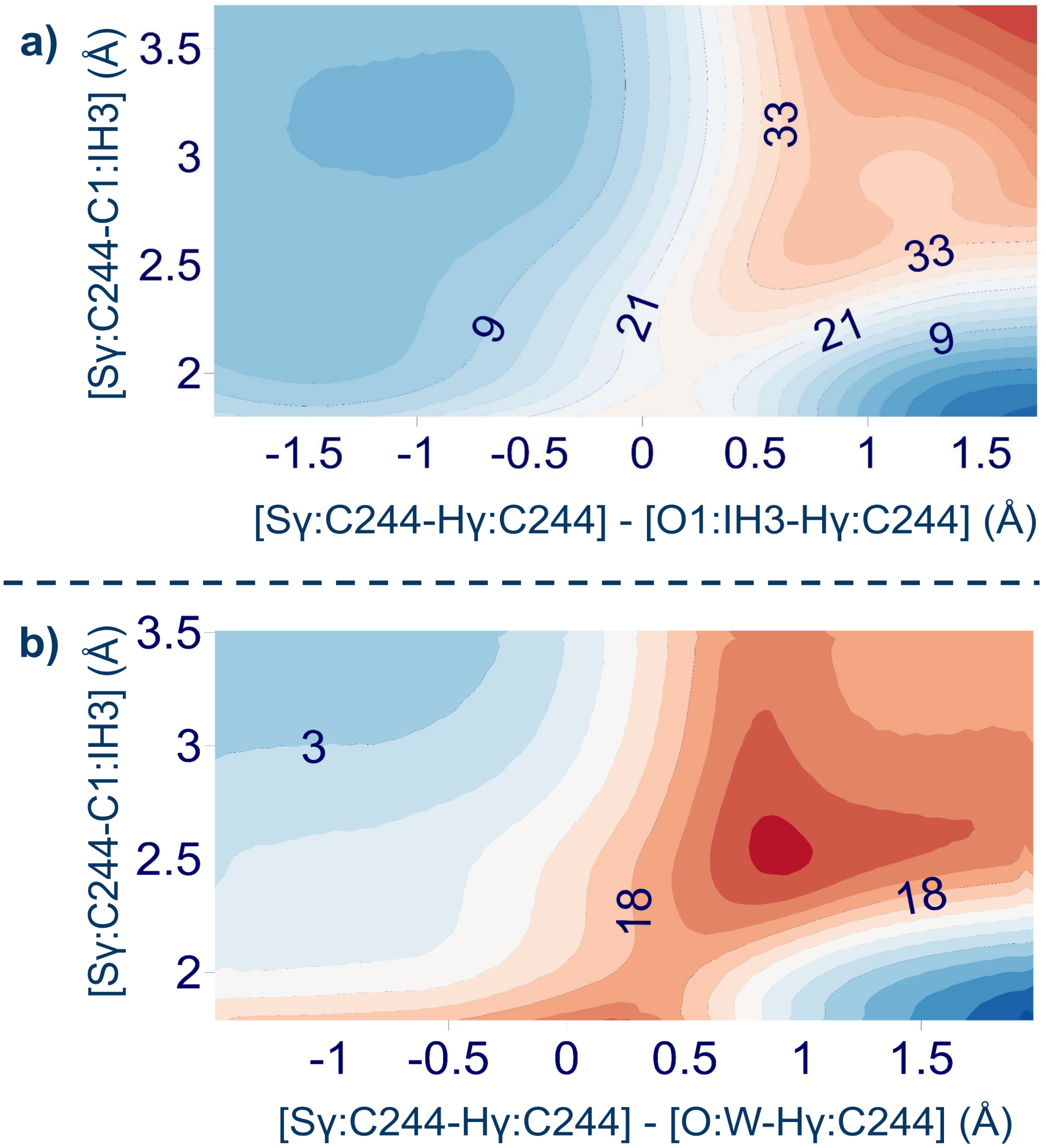


**Figure S7.** Free energy surfaces computed at PM6/MM level for the reaction mechanisms considered for the covalent binding inhibition of the RgpB gingipain. The values of energies are in kcal·mol$^{-1}$. Each level in the contour plots represent a 3 kcal·mol$^{-1}$ variation.

**IH3 MOL2 FILE:**

```
@<TRIPOS>MOLECULE
IH3
   52    53     1     0     0
SMALL
resp


@<TRIPOS>ATOM
      1 N1           1.9400    -1.3680     0.3020 n          1 IH3      -0.726019
      2 H1           1.9450    -1.7450    -0.6390 hn         1 IH3       0.333765
      3 C1           0.7530    -0.6430     0.6970 c3         1 IH3       0.212927
      4 H2           0.2950    -1.1350     1.5670 h1         1 IH3       0.102195
      5 C2           1.0970     0.8330     1.0800 c3         1 IH3      -0.429590
      6 C3          -0.2720    -0.6040    -0.4400 c          1 IH3       0.606860
      7 H3           0.1720     1.3430     1.3620 hc         1 IH3       0.131539
      8 H4           1.7950     0.8190     1.9200 hc         1 IH3       0.131539
      9 C4           1.7490     1.6100    -0.0840 c3         1 IH3       0.096711
     10 O1          -0.0310    -1.1430    -1.5030 o          1 IH3      -0.482681
     11 H5           2.6780     1.1100    -0.3770 hc         1 IH3       0.020502
     12 H6           1.0570     1.6330    -0.9370 hc         1 IH3       0.020502
     13 C5           2.0610     3.0440     0.3520 c3         1 IH3       0.229307
     14 H7           1.1340     3.5520     0.6470 h1         1 IH3       0.038436
     15 H8           2.7520     3.0300     1.2040 h1         1 IH3       0.038436
     16 N2           2.6850     3.7900    -0.7600 n          1 IH3      -0.732111
     17 H9           2.8220     3.2880    -1.6280 hn         1 IH3       0.372142
     18 C6           3.0770     5.0630    -0.7070 c          1 IH3       1.102047
     19 N3           2.9350     5.8020     0.3980 n          1 IH3      -1.038938
     20 N4           3.6300     5.6010    -1.8090 n          1 IH3      -1.038938
     21 H10          2.5220     5.4180     1.2360 hn         1 IH3       0.482854
```

```
    22 H11         3.2340    6.7660    0.4310 hn        1 IH3       0.482854
    23 H12         3.7480    5.0630   -2.6550 hn        1 IH3       0.482854
    24 H13         3.9480    6.5590   -1.8340 hn        1 IH3       0.482854
    25 C7          3.0180   -1.5550    1.1290 c         1 IH3       0.894639
    26 O2          3.0530   -1.1160    2.2750 o         1 IH3      -0.637783
    27 C8          4.1430   -2.3400    0.5090 ca        1 IH3      -0.187791
    28 C9          3.9510   -3.6800    0.1600 ca        1 IH3      -0.099678
    29 C10         5.3790   -1.7300    0.2810 ca        1 IH3      -0.099678
    30 C11         4.9940   -4.4100   -0.4150 ca        1 IH3      -0.125487
    31 H14         2.9890   -4.1540    0.3380 ha        1 IH3       0.125902
    32 C12         6.4210   -2.4600   -0.2940 ca        1 IH3      -0.125487
    33 H15         5.5290   -0.6880    0.5520 ha        1 IH3       0.125902
    34 C13         6.2290   -3.8000   -0.6430 ca        1 IH3      -0.099557
    35 H16         4.8450   -5.4520   -0.6850 ha        1 IH3       0.138002
    36 H17         7.0410   -4.3670   -1.0900 ha        1 IH3       0.137605
    37 H18         7.3820   -1.9850   -0.4720 ha        1 IH3       0.138002
    38 C14        -1.6110    0.1170   -0.2370 c3        1 IH3      -0.182194
    39 H19        -1.4880    0.9900    0.4160 h1        1 IH3       0.114107
    40 H20        -2.0140    0.4620   -1.1940 h1        1 IH3       0.114107
    41 O3         -2.4820   -0.8470    0.3630 os        1 IH3      -0.280252
    42 C15        -3.8240   -0.6280    0.1670 ca        1 IH3       0.114877
    43 C16        -4.5610    0.0780    1.1170 ca        1 IH3       0.118702
    44 C17        -4.4490   -1.1110   -0.9830 ca        1 IH3       0.118702
    45 C18        -5.9230    0.3020    0.9230 ca        1 IH3       0.269084
    46 C19        -5.8100   -0.8850   -1.1750 ca        1 IH3       0.269084
    47 C20        -6.5470   -0.1810   -0.2260 ca        1 IH3      -0.450119
    48 H21        -7.6060   -0.0080   -0.3800 ha        1 IH3       0.278665
    49 F1         -3.9580    0.5410    2.2220 f         1 IH3      -0.136356
    50 F2         -6.6280    0.9790    1.8360 f         1 IH3      -0.158344
    51 F3         -3.7450   -1.7890   -1.8960 f         1 IH3      -0.136356
    52 F4         -6.4120   -1.3470   -2.2760 f         1 IH3      -0.158344
@<TRIPOS>BOND
    1    1    2 1
```

```
2    1    3 1
3    1   25 1
4    3    4 1
5    3    5 1
6    3    6 1
7    5    7 1
8    5    8 1
9    5    9 1
10    6    10 2
11    6    38 1
12    9    11 1
13    9    12 1
14    9    13 1
15   13    14 1
16   13    15 1
17   13    16 1
18   16    17 1
19   16    18 2
20   18    19 1
21   18    20 1
22   19    21 1
23   19    22 1
24   20    23 1
25   20    24 1
26   25    26 2
27   25    27 1
28   27    28 ar
29   27    29 ar
30   28    30 ar
31   28    31 1
32   29    32 ar
33   29    33 1
34   30    34 ar
```

```
    35    30    35 1
    36    32    34 ar
    37    32    37 1
    38    34    36 1
    39    38    39 1
    40    38    40 1
    41    38    41 1
    42    41    42 1
    43    42    43 ar
    44    42    44 ar
    45    43    45 ar
    46    43    49 1
    47    44    46 ar
    48    44    51 1
    49    45    47 ar
    50    45    50 1
    51    46    47 ar
    52    46    52 1
    53    47    48 1
@<TRIPOS>SUBSTRUCTURE
     1 IH3         1 TEMP              0 ****  ****    0 ROOT
```

**IH3 FRCMOD FILE:**

```
Remark line goes here
MASS
n  14.010        0.530
hn 1.008         0.161
c3 12.010        0.878
h1 1.008         0.135
c  12.010        0.616
hc 1.008         0.135
o  16.000        0.434
ca 12.010        0.360
ha 1.008         0.135
os 16.000        0.465
f  19.000        0.320

BOND
hn-n   403.20   1.013
c3-n   328.70   1.462
c -n   427.60   1.379
c3-h1  330.60   1.097
c3-c3  300.90   1.538
c -c3  313.00   1.524
c3-hc  330.60   1.097
c -o   637.70   1.218
c -ca  345.90   1.491
ca-ca  461.10   1.398
ca-ha  345.80   1.086
c3-os  308.60   1.432
ca-os  376.60   1.370
ca-f   357.80   1.349
```

```
ANGLE
h1-c3-n    49.800     108.880
c3-c3-n    65.900     111.610
c -c3-n    67.000     109.060
n -c -o    74.200     123.050
ca-c -n    67.700     115.250
c3-n -hn   45.800     117.680
c -n -hn   48.300     117.550
c -n -c3   63.400     120.690
c3-c3-hc   46.300     109.800
c3-c3-c3   62.900     111.510
c3-c -o    67.400     123.200
c3-c -c3   62.000     116.500
c3-c3-h1   46.400     109.560
c -c3-h1   47.000     108.220
c -c3-c3   63.300     111.040
c -c3-os   68.000     109.210
hc-c3-hc   39.400     107.580
h1-c3-h1   39.200     108.460
n -c -n    72.900     113.560
hn-n -hn   39.600     117.950
c -ca-ca   64.300     120.330
ca-c -o    68.700     122.600
ca-ca-ca   66.600     120.020
ca-ca-ha   48.200     119.880
c3-os-ca   62.500     117.960
h1-c3-os   50.800     109.780
ca-ca-os   69.600     119.200
ca-ca-f    67.100     118.960

DIHE
hc-c3-c3-n    9    1.400         0.000           3.000
```

| | | | | |
|---|---|---|---|---|
| c3-c3-c3-n | 9 | 1.400 | 0.000 | 3.000 |
| o -c -c3-n | 6 | 0.000 | 180.000 | 2.000 |
| c3-c -c3-n | 6 | 0.000 | 180.000 | 2.000 |
| n -c -ca-ca | 4 | 4.000 | 180.000 | 2.000 |
| h1-c3-n -hn | 6 | 0.000 | 0.000 | 2.000 |
| c3-c3-n -hn | 6 | 0.000 | 0.000 | 2.000 |
| c -c3-n -hn | 6 | 0.000 | 0.000 | 2.000 |
| o -c -n -hn | 1 | 2.500 | 180.000 | -2.000 |
| o -c -n -hn | 1 | 2.000 | 0.000 | 1.000 |
| ca-c -n -hn | 4 | 10.000 | 180.000 | 2.000 |
| o -c -n -c3 | 4 | 10.000 | 180.000 | 2.000 |
| ca-c -n -c3 | 4 | 10.000 | 180.000 | 2.000 |
| c3-c3-c3-hc | 1 | 0.160 | 0.000 | 3.000 |
| c3-c3-c3-c3 | 1 | 0.180 | 0.000 | -3.000 |
| c3-c3-c3-c3 | 1 | 0.250 | 180.000 | -2.000 |
| c3-c3-c3-c3 | 1 | 0.200 | 180.000 | 1.000 |
| c3-c -c3-h1 | 6 | 0.000 | 180.000 | 2.000 |
| c3-c -c3-os | 6 | 0.000 | 180.000 | 2.000 |
| h1-c3-c3-hc | 9 | 1.400 | 0.000 | 3.000 |
| c3-c3-c3-h1 | 9 | 1.400 | 0.000 | 3.000 |
| o -c -c3-h1 | 1 | 0.800 | 0.000 | -1.000 |
| o -c -c3-h1 | 1 | 0.000 | 0.000 | -2.000 |
| o -c -c3-h1 | 1 | 0.080 | 180.000 | 3.000 |
| o -c -c3-c3 | 6 | 0.000 | 180.000 | 2.000 |
| c3-c -c3-c3 | 6 | 0.000 | 180.000 | 2.000 |
| c -c3-c3-hc | 9 | 1.400 | 0.000 | 3.000 |
| c -c3-c3-c3 | 9 | 1.400 | 0.000 | 3.000 |
| c -c3-os-ca | 3 | 1.150 | 0.000 | 3.000 |
| hc-c3-c3-hc | 1 | 0.150 | 0.000 | 3.000 |
| c3-c3-n -c | 1 | 0.500 | 180.000 | -4.000 |
| c3-c3-n -c | 1 | 0.150 | 180.000 | -3.000 |
| c3-c3-n -c | 1 | 0.000 | 0.000 | -2.000 |
| c3-c3-n -c | 1 | 0.530 | 0.000 | 1.000 |

```
o -c -c3-os   6    0.000       180.000           2.000
n -c -n -c3   4   10.000       180.000           2.000
h1-c3-n -c    6    0.000         0.000           2.000
n -c -n -hn   4   10.000       180.000           2.000
c -c3-n -c    1    0.850       180.000          -2.000
c -c3-n -c    1    0.800         0.000           1.000
c -ca-ca-ca   4   14.500       180.000           2.000
c -ca-ca-ha   4   14.500       180.000           2.000
o -c -ca-ca   4    4.000       180.000           2.000
ca-ca-ca-ca   4   14.500       180.000           2.000
ca-ca-ca-ha   4   14.500       180.000           2.000
ha-ca-ca-ha   4   14.500       180.000           2.000
ca-ca-os-c3   2    1.800       180.000           2.000
h1-c3-os-ca   3    1.150         0.000           3.000
ca-ca-ca-os   4   14.500       180.000           2.000
f -ca-ca-os   4   14.500       180.000           2.000
ca-ca-ca-f    4   14.500       180.000           2.000
f -ca-ca-f    4   14.500       180.000           2.000
f -ca-ca-ha   4   14.500       180.000           2.000

IMPROPER
c -c3-n -hn         1.1          180.0         2.0
c3-c3-c -o         10.5          180.0         2.0          Using general improper torsional angle  X- X- c- o,
penalty score=  6.0)
n -n -c -n          1.1          180.0         2.0          Using the default value
c -hn-n -hn         1.1          180.0         2.0          Using general improper torsional angle  X- X- n-hn,
penalty score=  6.0)
ca-n -c -o         10.5          180.0         2.0          Using general improper torsional angle  X- X- c- o,
penalty score=  6.0)
c -ca-ca-ca         1.1          180.0         2.0          Using the default value
ca-ca-ca-ha         1.1          180.0         2.0          Using general improper torsional angle  X- X-ca-ha,
penalty score=  6.0)
ca-ca-ca-os         1.1          180.0         2.0          Using the default value
```

```
ca-ca-ca-f          1.1          180.0          2.0

NONBON
  n          1.8240  0.1700
  hn          0.6000  0.0157
  c3          1.9080  0.1094
  h1          1.3870  0.0157
  c          1.9080  0.0860
  hc          1.4870  0.0157
  o          1.6612  0.2100
  ca          1.9080  0.0860
  ha          1.4590  0.0150
  os          1.6837  0.1700
  f          1.7500  0.0610
```

**IHV MOL2 FILE:**

```
@<TRIPOS>MOLECULE
IHV
   28    27     1     0     0
SMALL
resp



@<TRIPOS>ATOM
      1 N1           2.8560     0.3100    -0.6880 n          1 IHV      -0.582456
      2 H1           3.0590    -0.3380    -1.4410 hn          1 IHV       0.325682
      3 C1           2.5340    -0.2770     0.5980 c3          1 IHV       0.263671
      4 H2           3.2640     0.0610     1.3490 h1          1 IHV       0.106044
      5 C2           1.1090     0.1440     1.0650 c3          1 IHV      -0.564600
      6 C3           2.5720    -1.7900     0.5320 c          1 IHV       0.320266
      7 H3           0.9090    -0.3310     2.0340 hc          1 IHV       0.162640
      8 H4           1.0850     1.2310     1.1580 hc          1 IHV       0.162640
      9 C4           0.0020    -0.2840     0.0770 c3          1 IHV       0.184094
     10 O1           2.8390    -2.4110    -0.4730 o          1 IHV      -0.436977
     11 H5           0.1860     0.1820    -0.8970 hc          1 IHV       0.026178
     12 H6           0.0120    -1.3760    -0.0290 hc          1 IHV       0.026178
     13 C5          -1.3650     0.1620     0.6010 c3          1 IHV       0.187846
     14 H7          -1.5580    -0.3010     1.5770 h1          1 IHV       0.047366
     15 H8          -1.3840     1.2530     0.7090 h1          1 IHV       0.047366
     16 N2          -2.4280    -0.2450    -0.3410 n          1 IHV      -0.736190
     17 H9          -2.1390    -0.7420    -1.1730 hn          1 IHV       0.355037
     18 C6          -3.7310    -0.0070    -0.1770 c          1 IHV       1.153561
     19 N3          -4.2040     0.6420     0.8910 n          1 IHV      -1.089551
     20 N4          -4.5770    -0.4430    -1.1280 n          1 IHV      -1.089551
     21 H10         -3.5860     0.9780     1.6160 hn          1 IHV       0.500425
```

```
     22 H11        -5.1920     0.8180     1.0070 hn        1 IHV       0.500425
     23 H12        -4.2440    -0.9360    -1.9440 hn        1 IHV       0.500425
     24 H13        -5.5740    -0.2960    -1.0640 hn        1 IHV       0.500425
     25 C7          2.8950     1.6540    -0.9090 c         1 IHV       0.568545
     26 O2          2.6660     2.5030    -0.0630 o         1 IHV      -0.562933
     27 H14         2.3340    -2.3010     1.4870 h4        1 IHV       0.076394
     28 H15         3.1600     1.9000    -1.9540 h5        1 IHV       0.047051
@<TRIPOS>BOND
     1     1     2 1
     2     1     3 1
     3     1    25 1
     4     3     4 1
     5     3     5 1
     6     3     6 1
     7     5     7 1
     8     5     8 1
     9     5     9 1
    10     6    10 2
    11     6    27 1
    12     9    11 1
    13     9    12 1
    14     9    13 1
    15    13    14 1
    16    13    15 1
    17    13    16 1
    18    16    17 1
    19    16    18 2
    20    18    19 1
    21    18    20 1
    22    19    21 1
    23    19    22 1
    24    20    23 1
    25    20    24 1
```

```
   26    25    26 2
   27    25    28 1
@<TRIPOS>SUBSTRUCTURE
     1 IHV         1 TEMP              0 ****  ****    0 ROOT
```

**IHV FRCMOD FILE:**

```
Remark line goes here
MASS
n  14.010        0.530
hn 1.008         0.161
c3 12.010        0.878
h1 1.008         0.135
c  12.010        0.616
hc 1.008         0.135
o  16.000        0.434
nh 14.010        0.530
cz 12.010        0.360
h4 1.008         0.135
h5 1.008         0.135

BOND
hn-n   403.20   1.013
c3-n   328.70   1.462
c -n   427.60   1.379
c3-h1  330.60   1.097
c3-c3  300.90   1.538
c -c3  313.00   1.524
c3-hc  330.60   1.097
c -o   637.70   1.218
c -h4  310.70   1.112
c3-nh  326.60   1.464
hn-nh  404.60   1.012
cz-nh  488.00   1.339
c -h5  319.70   1.105
```

```
ANGLE
h1-c3-n    49.800     108.880
c3-c3-n    65.900     111.610
c -c3-n    67.000     109.060
n -c -o    74.200     123.050
h5-c -n    51.400     112.160
c3-n -hn   45.800     117.680
c -n -hn   48.300     117.550
c -n -c3   63.400     120.690
c3-c3-hc   46.300     109.800
c3-c3-c3   62.900     111.510
c3-c -o    67.400     123.200
c3-c -h4   45.600     114.640
c3-c3-h1   46.400     109.560
c -c3-h1   47.000     108.220
c -c3-c3   63.300     111.040
hc-c3-hc   39.400     107.580
c3-c3-nh   66.200     110.460
h4-c -o    54.200     120.700
c3-nh-hn   46.100     115.990
c3-nh-cz   62.900     125.460
h1-c3-h1   39.200     108.460
h1-c3-nh   49.600     109.790
nh-cz-nh   73.000     120.140
cz-nh-hn   48.800     121.150
hn-nh-hn   40.100     115.120
h5-c -o    53.700     123.650

DIHE
hc-c3-c3-n    9    1.400        0.000          3.000
c3-c3-c3-n    9    1.400        0.000          3.000
o -c -c3-n    6    0.000      180.000          2.000
h4-c -c3-n    6    0.000      180.000          2.000
```

```
h1-c3-n -hn   6    0.000         0.000           2.000
c3-c3-n -hn   6    0.000         0.000           2.000
c -c3-n -hn   6    0.000         0.000           2.000
o -c -n -hn   1    2.500       180.000          -2.000
o -c -n -hn   1    2.000         0.000           1.000
h5-c -n -hn   4   10.000       180.000           2.000
o -c -n -c3   4   10.000       180.000           2.000
h5-c -n -c3   4   10.000       180.000           2.000
c3-c3-c3-hc   1    0.160         0.000           3.000
c3-c3-c3-c3   1    0.180         0.000          -3.000
c3-c3-c3-c3   1    0.250       180.000          -2.000
c3-c3-c3-c3   1    0.200       180.000           1.000
h1-c3-c3-hc   9    1.400         0.000           3.000
c3-c3-c3-h1   9    1.400         0.000           3.000
o -c -c3-h1   1    0.800         0.000          -1.000
o -c -c3-h1   1    0.000         0.000          -2.000
o -c -c3-h1   1    0.080       180.000           3.000
h4-c -c3-h1   6    0.000       180.000           2.000
o -c -c3-c3   6    0.000       180.000           2.000
h4-c -c3-c3   6    0.000       180.000           2.000
c3-c3-c3-nh   9    1.400         0.000           3.000
c -c3-c3-hc   9    1.400         0.000           3.000
c -c3-c3-c3   9    1.400         0.000           3.000
hc-c3-c3-hc   1    0.150         0.000           3.000
c3-c3-nh-hn   6    0.000         0.000           2.000
c3-c3-nh-cz   6    0.000         0.000           2.000
hc-c3-c3-nh   9    1.400         0.000           3.000
nh-cz-nh-c3   4    2.700       180.000           2.000      same as X -c2-nh-X , penalty score=462.5
h1-c3-nh-hn   6    0.000         0.000           2.000
h1-c3-nh-cz   6    0.000         0.000           2.000
nh-cz-nh-hn   4    2.700       180.000           2.000      same as X -c2-nh-X , penalty score=462.5
h1-c3-n -c    6    0.000         0.000           2.000
c3-c3-n -c    1    0.500       180.000          -4.000
```

```
c3-c3-n -c    1    0.150       180.000          -3.000
c3-c3-n -c    1    0.000         0.000          -2.000
c3-c3-n -c    1    0.530         0.000           1.000
c -c3-n -c    1    0.850       180.000          -2.000
c -c3-n -c    1    0.800         0.000           1.000

IMPROPER
c -c3-n -hn         1.1          180.0         2.0
c3-h4-c -o         10.5          180.0         2.0          Using general improper torsional angle  X- X- c- o,
penalty score=  6.0)
c3-cz-nh-hn         1.1          180.0         2.0          Using the default value
nh-nh-cz-nh         1.1          180.0         2.0          Using the default value
cz-hn-nh-hn         1.1          180.0         2.0          Same as X -X -na-hn, penalty score= 41.2 (use general
term))
h5-n -c -o         10.5          180.0         2.0          Using general improper torsional angle  X- X- c- o,
penalty score=  6.0)

NONBON
  n          1.8240  0.1700
  hn         0.6000  0.0157
  c3         1.9080  0.1094
  h1         1.3870  0.0157
  c          1.9080  0.0860
  hc         1.4870  0.0157
  o          1.6612  0.2100
  nh         1.8240  0.1700
  cz         1.9080  0.0860
  h4         1.4090  0.0150
  h5         1.3590  0.0150
```

**QM REGION COORDINATES FOR TRANSITION STATE (TS$_W$) STRUCTURE:**

96

| | | | |
|---|---|---|---|
| C | 47.624249 | 81.588348 | 62.890717 |
| H | 47.985661 | 80.606384 | 62.530949 |
| H | 47.235641 | 82.132645 | 62.014744 |
| C | 48.876663 | 82.286835 | 63.496735 |
| O | 49.399414 | 83.201538 | 62.739300 |
| O | 49.314838 | 81.875092 | 64.590317 |
| C | 52.514141 | 84.329651 | 57.975819 |
| O | 52.967987 | 85.315399 | 58.591171 |
| N | 53.189388 | 83.172523 | 57.843319 |
| H | 52.873268 | 82.488045 | 57.151871 |
| C | 54.491104 | 82.899460 | 58.443508 |
| H | 54.989918 | 83.868607 | 58.619629 |
| C | 54.297817 | 82.200020 | 59.826122 |
| H | 53.805382 | 81.227097 | 59.650463 |
| H | 55.299221 | 81.993080 | 60.239849 |
| C | 53.492695 | 82.987503 | 60.825127 |
| N | 53.986176 | 84.101418 | 61.496586 |
| C | 52.992359 | 84.463181 | 62.329380 |
| H | 53.065025 | 85.285957 | 63.041138 |
| N | 51.894096 | 83.658714 | 62.229549 |
| H | 50.889011 | 83.583122 | 62.714272 |
| C | 52.205009 | 82.722893 | 61.274761 |
| H | 51.488773 | 81.939369 | 61.045029 |
| C | 55.292084 | 82.085571 | 57.379807 |
| O | 54.838585 | 81.932480 | 56.241741 |
| N | 56.488247 | 81.529961 | 57.772121 |
| H | 56.864220 | 81.890640 | 58.652100 |

|  |  |  |  |
|---|---|---|---|
| C | 57.319881 | 80.719765 | 56.840694 |
| H | 56.693230 | 80.574890 | 55.948055 |
| H | 58.205940 | 81.297165 | 56.519791 |
| C | 58.588017 | 85.155396 | 55.607357 |
| O | 59.435051 | 85.732475 | 54.920357 |
| N | 58.942921 | 84.408653 | 56.717377 |
| H | 58.197227 | 83.967979 | 57.265423 |
| C | 60.357407 | 84.137390 | 57.129753 |
| H | 60.861404 | 85.097099 | 57.326214 |
| C | 60.292995 | 83.269890 | 58.418858 |
| H | 59.416660 | 82.592407 | 58.334961 |
| H | 61.180019 | 82.617241 | 58.468620 |
| S | 60.252052 | 84.013885 | 60.125420 |
| H | 59.858013 | 82.354027 | 60.579681 |
| N | 58.031010 | 86.013763 | 61.984367 |
| H | 59.043861 | 85.911034 | 61.941349 |
| C | 57.292057 | 85.265121 | 60.975861 |
| H | 56.249912 | 85.136116 | 61.362591 |
| C | 57.243580 | 86.039726 | 59.650162 |
| C | 57.836182 | 83.837532 | 60.972218 |
| H | 58.277637 | 86.245583 | 59.323738 |
| H | 56.779430 | 85.413284 | 58.865749 |
| C | 56.456009 | 87.359116 | 59.819996 |
| O | 57.412155 | 82.955727 | 60.106564 |
| H | 56.712597 | 87.810471 | 60.794395 |
| H | 56.756618 | 88.077858 | 59.035427 |
| C | 54.952545 | 87.092468 | 59.741367 |
| H | 54.689339 | 86.712021 | 58.740952 |
| H | 54.648239 | 86.315094 | 60.466152 |
| N | 54.154266 | 88.298431 | 60.026821 |
| H | 54.546959 | 89.205368 | 59.773037 |
| C | 52.866600 | 88.282593 | 60.410492 |
| N | 52.138428 | 87.168312 | 60.364517 |

| | | | |
|---|---|---|---|
| N | 52.299870 | 89.462830 | 60.815521 |
| H | 51.319454 | 87.057076 | 60.959938 |
| H | 52.414272 | 86.360451 | 59.753647 |
| H | 52.864910 | 90.090050 | 61.386295 |
| H | 51.309010 | 89.449898 | 61.048088 |
| C | 57.461437 | 86.522888 | 63.129932 |
| O | 56.232346 | 86.653954 | 63.245193 |
| C | 58.413204 | 86.914581 | 64.236282 |
| C | 59.809631 | 86.719696 | 64.203552 |
| C | 57.838734 | 87.461891 | 65.398819 |
| C | 60.607372 | 87.088608 | 65.292809 |
| H | 60.304302 | 86.239594 | 63.353630 |
| C | 58.634380 | 87.837006 | 66.485268 |
| H | 56.752354 | 87.552536 | 65.433685 |
| C | 60.024380 | 87.656265 | 66.433800 |
| H | 61.684834 | 86.901726 | 65.255653 |
| H | 60.650082 | 87.942924 | 67.284973 |
| H | 58.167557 | 88.259155 | 67.380859 |
| C | 58.002571 | 83.278465 | 62.400146 |
| H | 57.095287 | 83.608543 | 62.935635 |
| H | 58.013069 | 82.177238 | 62.386284 |
| O | 59.192764 | 83.753822 | 63.049622 |
| C | 59.350502 | 83.567299 | 64.393921 |
| C | 58.357269 | 83.305252 | 65.359756 |
| C | 60.671997 | 83.749725 | 64.856606 |
| C | 58.699554 | 83.222214 | 66.719734 |
| C | 60.986938 | 83.680016 | 66.213646 |
| C | 60.006706 | 83.402115 | 67.167931 |
| H | 60.249889 | 83.341789 | 68.229630 |
| F | 57.054165 | 83.147308 | 65.027184 |
| F | 57.713142 | 82.967545 | 67.599609 |
| F | 61.627850 | 84.079956 | 63.965172 |
| F | 62.264027 | 83.899521 | 66.587868 |

| | | | |
|---|---|---|---|
| O | 59.205906 | 81.383141 | 60.611393 |
| H | 59.493832 | 80.827522 | 59.856094 |
| H | 58.213539 | 81.966354 | 60.326187 |

## QM REGION COORDINATES FOR REACTANTS (R) STRUCTURE:

96

| | | | |
|---|---|---|---|
| C | 47.619030 | 81.652100 | 62.917931 |
| H | 47.930462 | 80.691505 | 62.486317 |
| H | 47.238106 | 82.266220 | 62.101234 |
| C | 48.854809 | 82.261345 | 63.563835 |
| O | 49.458378 | 83.169693 | 62.877743 |
| O | 49.251507 | 81.852386 | 64.652275 |
| C | 52.554493 | 84.313766 | 57.990585 |
| O | 53.134624 | 85.319061 | 58.422516 |
| N | 53.172291 | 83.085320 | 58.033230 |
| H | 52.744503 | 82.256271 | 57.617256 |
| C | 54.531349 | 82.902992 | 58.591866 |
| H | 55.006588 | 83.911484 | 58.775005 |
| C | 54.425426 | 82.154243 | 59.938068 |
| H | 54.000816 | 81.139549 | 59.783466 |
| H | 55.444881 | 82.000282 | 60.351742 |
| C | 53.595566 | 82.894234 | 60.933346 |
| N | 54.031582 | 84.024338 | 61.627106 |
| C | 53.003342 | 84.335754 | 62.490551 |
| H | 53.038746 | 85.136429 | 63.210556 |
| N | 51.934547 | 83.459236 | 62.358765 |
| H | 50.885906 | 83.393555 | 62.900272 |
| C | 52.300247 | 82.556084 | 61.372246 |
| H | 51.655319 | 81.753777 | 61.080029 |
| C | 55.341171 | 82.159569 | 57.505238 |

| | | | |
|---|---|---|---|
| O | 54.905567 | 82.002022 | 56.376041 |
| N | 56.593189 | 81.650124 | 57.825058 |
| H | 57.067860 | 81.916367 | 58.689701 |
| C | 57.325733 | 80.800293 | 56.840263 |
| H | 56.681171 | 80.628952 | 55.950943 |
| H | 58.232609 | 81.317307 | 56.486176 |
| C | 58.580002 | 85.116982 | 55.811672 |
| O | 59.478336 | 85.650558 | 55.170868 |
| N | 58.938339 | 84.376205 | 56.953995 |
| H | 58.250309 | 84.006187 | 57.609718 |
| C | 60.394135 | 84.142677 | 57.262081 |
| H | 60.889664 | 85.135872 | 57.374752 |
| C | 60.564884 | 83.283676 | 58.518623 |
| H | 60.039143 | 82.309006 | 58.437901 |
| H | 61.641384 | 83.066490 | 58.680569 |
| S | 59.977577 | 84.121994 | 60.026093 |
| H | 60.106544 | 83.089928 | 60.911839 |
| N | 58.152016 | 85.928185 | 62.274651 |
| H | 59.118172 | 85.600319 | 62.247219 |
| C | 57.265636 | 85.491150 | 61.181919 |
| H | 56.177200 | 85.644661 | 61.517956 |
| C | 57.510731 | 86.287369 | 59.885815 |
| C | 57.393673 | 83.981606 | 60.984425 |
| H | 58.567341 | 86.624252 | 59.826946 |
| H | 57.374828 | 85.629143 | 59.002033 |
| C | 56.576904 | 87.499908 | 59.812447 |
| O | 57.185608 | 83.436493 | 59.917519 |
| H | 56.563446 | 88.005440 | 60.806480 |
| H | 56.979748 | 88.245552 | 59.102554 |
| C | 55.163853 | 87.055634 | 59.413750 |
| H | 55.089539 | 86.779892 | 58.333702 |
| H | 54.869896 | 86.128983 | 59.976357 |
| N | 54.204926 | 88.163948 | 59.690037 |

| | | | |
|---|---|---|---|
| H | 54.496338 | 89.100258 | 59.409351 |
| C | 52.957371 | 87.968269 | 60.226891 |
| N | 52.491096 | 86.738342 | 60.585888 |
| N | 52.138123 | 89.099785 | 60.403099 |
| H | 51.547501 | 86.598198 | 60.932854 |
| H | 52.974819 | 85.869293 | 60.260418 |
| H | 52.578228 | 90.005951 | 60.523102 |
| H | 51.288803 | 88.996475 | 60.952667 |
| C | 57.665821 | 86.743706 | 63.300457 |
| O | 56.516346 | 87.160828 | 63.234066 |
| C | 58.598293 | 87.050323 | 64.428070 |
| C | 59.984734 | 86.841019 | 64.372543 |
| C | 58.026306 | 87.590256 | 65.591118 |
| C | 60.787197 | 87.171761 | 65.466377 |
| H | 60.454578 | 86.431007 | 63.477837 |
| C | 58.832932 | 87.932838 | 66.678398 |
| H | 56.945030 | 87.734337 | 65.640907 |
| C | 60.213360 | 87.723663 | 66.616966 |
| H | 61.862041 | 87.001968 | 65.423752 |
| H | 60.840851 | 87.993660 | 67.465263 |
| H | 58.383060 | 88.360291 | 67.574249 |
| C | 57.543045 | 83.127609 | 62.251957 |
| H | 56.686733 | 83.358826 | 62.916035 |
| H | 57.577499 | 82.044701 | 62.029530 |
| O | 58.845268 | 83.441170 | 62.825714 |
| C | 59.025730 | 83.402893 | 64.181908 |
| C | 58.110134 | 83.183212 | 65.224686 |
| C | 60.380009 | 83.638023 | 64.539268 |
| C | 58.544132 | 83.227760 | 66.564110 |
| C | 60.784962 | 83.653351 | 65.875420 |
| C | 59.872490 | 83.455170 | 66.918098 |
| H | 60.189770 | 83.475769 | 67.963387 |
| F | 56.826027 | 82.920624 | 64.998749 |

| | | | |
|---|---|---|---|
| F | 57.650631 | 83.036926 | 67.525543 |
| F | 61.257942 | 83.826042 | 63.559460 |
| F | 62.062740 | 83.862968 | 66.168648 |
| O | 59.020229 | 81.447556 | 60.360493 |
| H | 59.034721 | 80.501060 | 60.484417 |
| H | 58.182602 | 81.675430 | 59.930988 |

## QM REGION COORDINATES FOR PRODUCT (P) STRUCTURE:

96

| | | | |
|---|---|---|---|
| C | 47.618038 | 81.639740 | 62.923149 |
| H | 47.946659 | 80.676514 | 62.511894 |
| H | 47.238941 | 82.239136 | 62.094959 |
| C | 48.827061 | 82.274765 | 63.587093 |
| O | 49.391190 | 83.219872 | 62.919884 |
| O | 49.242867 | 81.855927 | 64.665451 |
| C | 52.527370 | 84.318550 | 58.035431 |
| O | 53.130333 | 85.326347 | 58.419315 |
| N | 53.084499 | 83.068665 | 58.198254 |
| H | 52.606453 | 82.227005 | 57.879925 |
| C | 54.445545 | 82.882973 | 58.741196 |
| H | 54.930874 | 83.895096 | 58.897675 |
| C | 54.373547 | 82.167603 | 60.105595 |
| H | 53.981419 | 81.136208 | 59.986385 |
| H | 55.399391 | 82.070061 | 60.522785 |
| C | 53.528046 | 82.910553 | 61.086201 |
| N | 53.972340 | 84.031166 | 61.789093 |
| C | 52.946278 | 84.349831 | 62.648632 |
| H | 52.987690 | 85.142563 | 63.376839 |
| N | 51.866165 | 83.489258 | 62.499241 |
| H | 50.814831 | 83.447929 | 63.018196 |

| | | | |
|---|---|---|---|
| C | 52.223640 | 82.587090 | 61.509007 |
| H | 51.565517 | 81.801254 | 61.199379 |
| C | 55.230415 | 82.155380 | 57.626453 |
| O | 54.886082 | 82.187569 | 56.457985 |
| N | 56.394520 | 81.472015 | 57.960712 |
| H | 56.703934 | 81.399399 | 58.929409 |
| C | 57.175476 | 80.753708 | 56.907406 |
| H | 56.530792 | 80.578033 | 56.022045 |
| H | 58.021664 | 81.375763 | 56.571800 |
| C | 58.577030 | 85.134399 | 55.785793 |
| O | 59.479023 | 85.644379 | 55.133453 |
| N | 58.912357 | 84.405533 | 56.937733 |
| H | 58.200169 | 84.110146 | 57.613831 |
| C | 60.352734 | 84.138344 | 57.287708 |
| H | 60.875732 | 85.113022 | 57.431206 |
| C | 60.353416 | 83.308266 | 58.589130 |
| H | 59.685753 | 82.425400 | 58.515045 |
| H | 61.379120 | 82.929153 | 58.788349 |
| S | 59.997543 | 84.195732 | 60.162758 |
| H | 59.607273 | 81.454323 | 60.873539 |
| N | 58.334320 | 85.974419 | 62.145943 |
| H | 59.341972 | 85.798340 | 62.187492 |
| C | 57.627186 | 85.487236 | 60.939724 |
| H | 56.523182 | 85.351830 | 61.205536 |
| C | 57.719635 | 86.448700 | 59.745930 |
| C | 58.206066 | 84.073830 | 60.651325 |
| H | 58.737560 | 86.879906 | 59.663620 |
| H | 57.566132 | 85.883324 | 58.803284 |
| C | 56.684608 | 87.574867 | 59.869183 |
| O | 57.502934 | 83.520950 | 59.566433 |
| H | 56.814228 | 88.081329 | 60.853638 |
| H | 56.874718 | 88.341789 | 59.096359 |
| C | 55.262875 | 87.006477 | 59.763504 |

| | | | |
|---|---|---|---|
| H | 55.074162 | 86.509033 | 58.778816 |
| H | 55.119400 | 86.221931 | 60.562210 |
| N | 54.272537 | 88.106606 | 59.951012 |
| H | 54.568531 | 89.045822 | 59.686600 |
| C | 52.976685 | 87.882858 | 60.339718 |
| N | 52.504604 | 86.642426 | 60.651810 |
| N | 52.115894 | 88.996841 | 60.421253 |
| H | 51.538139 | 86.483124 | 60.917274 |
| H | 53.025913 | 85.786026 | 60.366444 |
| H | 52.521786 | 89.908035 | 60.607151 |
| H | 51.214500 | 88.869377 | 60.874767 |
| C | 57.702698 | 86.685707 | 63.173000 |
| O | 56.531475 | 87.018417 | 63.063381 |
| C | 58.553169 | 87.008026 | 64.366692 |
| C | 59.950386 | 86.868668 | 64.386467 |
| C | 57.897926 | 87.494148 | 65.508347 |
| C | 60.679604 | 87.234619 | 65.519142 |
| H | 60.486839 | 86.479614 | 63.519054 |
| C | 58.629639 | 87.870522 | 66.637970 |
| H | 56.809498 | 87.566032 | 65.509605 |
| C | 60.021416 | 87.746246 | 66.643257 |
| H | 61.763523 | 87.124748 | 65.526749 |
| H | 60.593479 | 88.045135 | 67.521400 |
| H | 58.110828 | 88.257912 | 67.514549 |
| C | 57.993183 | 83.220192 | 61.923492 |
| H | 57.114536 | 83.599632 | 62.485889 |
| H | 57.823799 | 82.143776 | 61.726776 |
| O | 59.233646 | 83.323936 | 62.679924 |
| C | 59.191051 | 83.377937 | 64.047760 |
| C | 58.132179 | 83.153931 | 64.946121 |
| C | 60.464268 | 83.672989 | 64.602501 |
| C | 58.346203 | 83.276321 | 66.333206 |
| C | 60.654205 | 83.753662 | 65.983322 |

|  |  |  |  |
|---|---|---|---|
| C | 59.594818 | 83.569656 | 66.879105 |
| H | 59.741795 | 83.649231 | 67.958458 |
| F | 56.915462 | 82.806969 | 64.534943 |
| F | 57.320648 | 83.091530 | 67.154282 |
| F | 61.478100 | 83.852341 | 63.762718 |
| F | 61.865356 | 84.016724 | 66.460701 |
| O | 58.822933 | 81.131874 | 60.400723 |
| H | 58.874947 | 80.177597 | 60.390125 |
| H | 57.759945 | 82.549774 | 59.460674 |

# ANNEX C. ARTICLE 3

# Molecular Basis of RNA-Driven ATP Hydrolysis in DExH-box Helicases

Santiago Movilla[1], Maite Roca[1], Vicent Moliner[1] and Alessandra Magistrato[2,*]

[1]BioComp Group, Institute of Advanced Materials (INAM), Universitat Jaume I, 12071, Castellón, Spain.

[2]Department National Research Council of Italy (CNR) - Institute of Material (IOM) c/o International School for Advanced Studies (SISSA), Trieste, Italy

---

**ABSTRACT:** The spliceosome machinery catalyzes precursor-messenger (pre-m)RNA splicing. In each cycle, the spliceosome experiences massive compositional and conformational remodeling fueled by the concerted action of specific RNA-dependent ATPases/helicases. Intriguingly, these enzymes are allosterically activated to perform ATP hydrolysis and trigger helicase activity only upon pre-mRNA binding. Yet, the molecular mechanism underlying the RNA-driven regulation of their ATPase function remains elusive.

Here, we focus on the Prp2 ATPase/helicase which contributes to reshaping the spliceosome into its catalytic-competent state. By performing classical and quantum-classical molecular dynamics simulations, we unprecedentedly unlock the molecular terms governing Prp2 ATPase/helicase function. Namely, we dissect the molecular mechanism of ATP hydrolysis and we disclose that RNA binding allosterically triggers the formation of a set of interactions linking the RNA binding tunnel to the catalytic site. This activates Prp2's ATPase function by optimally placing the nucleophilic water and general base of the enzymatic process to perform ATP hydrolysis. The key structural motifs, mechanically coupling RNA gripping and ATPase/helicase function, are conserved across all DExH-box helicases. This mechanism could thus be broadly applicable to all DExH-box helicases family.

---

## 1. INTRODUCTION

Premature-messenger RNA (pre-mRNA) splicing denotes the removal of non-coding sequences (introns) from coding tracts (exons) to obtain a protein coding mRNA filament and functional non coding RNAs strands.[1–3] The splicing labor is carried out by a complex biomolecular machinery called spliceosome. This stunning macromolecular engine is composed of five small nuclear ribonucleoproteins ((snRNP), i.e. U1, U2, U4/U6 and U5), and dozens of other associated protein cofactors. For each intron to be excised, the spliceosome undergoes assembly, activation, catalysis, and disassembly with an impressive molecular accuracy.[4–6] Splicing fidelity indeed requires a sequence of stages to be rigorously followed at each splicing cycle.

First, pre-mRNA is recognized by the U1 snRNP, forming the E complex (E). Next, the latter associates with U2 snRNP to constitute the prespliceosome A complex. Then, in the following part of the cycle, the system is remodeled in three stages to obtain a catalytic competent spliceosome complex.[7] Namely, from pre-B the spliceosome heads to the precatalytic spliceosome (B), to the activated spliceosome (Bact) and to the catalytically activated spliceosome (B*). The first splicing (branching) step occurs in this catalytically-competent complex.[8] Then the spliceosome, evolves to complex (C) and step II activated complex (C*) where the second (exon ligation) step takes place.[9] The postcatalytic complex (P) assembles after the exon ligation catalytic step finally gives way to the intron lariat spliceosome (ILS), terminating the splicing cycle.[10,11]

All of the aforementioned steps require the concerted action of eight conserved ATPase/helicases, which fuel the complex conformational and compositional remodeling of the spliceosome through the hydrolysis of ATP.[12–16] These helicases belong to the DEAH/D-box (Prp2, Prp22, Prp43, Prp16 and Prp5, Sub2, Prp28, respectively) and Ski2-like (Brr2) families.[12–14,17] Remarkably, some spliceosomal helicases perform pre-mRNA proofreading (i.e. they promote ATP hydrolysis only upon binding of optimal client RNA substrates), thus increasing splicing fidelity.[18] Nevertheless, their proofreading activity is only observed *in vivo*. Whereas *in vitro*, single proteins are found to be sequence-independent. This is possibly because the proofreading activity is exerted by protein-protein-RNA interactions formed within the spliceosome complex, which act in concert with helicases.[18,19]

The DEAH-box ATPase/helicase Prp2 modulate the spliceosome transition to its crucial catalytically competent state, the B* complex. Prp2 complexes the pre-mRNA and hydrolyzes ATP to perform the stepwise translocation of pre-mRNA towards its 3' end.[20–23] Prp2's architecture is composed by two RecA-like domains, with a b-hairpin in the RecA2 domain, a winged helix (WH), a helix-bundle domain (HB) and an oligosaccharide-binding fold (OB) subdomain.[13,24] This core structure is conserved among other spliceosomal DEAH-box ATPases (i.e. Prp16, Prp22 and Prp43). In all these enzymes, a single strand RNA, upon binding in a tunnel formed at the interface of the RecA1, RecA2 and the HB, OB, WH domains, allosterically activates the ATPase function. Recent Cryo-EM and X-ray structures solved for yeast *C. Cerevisiae* and fungi *C. Thermophilum* Prp2 in presence of RNA, ADP/ATP, isolated or in within the spliceosome, supplied key information to formulate mechanistic hypothesis on the Prp2 helicase activity.[19] Namely, these structures suggested that the alternation of ATP, ADP-bound and APO states may induce pre-mRNA translocation at the rate of one RNA base per each ATP molecule hydrolyzed.[19] Nevertheless, the molecular mechanism of Prp2's ATP hydrolysis and of its RNA-driven allosteric regulation remains elusive. Yet, a complete mechanistic understanding of pre-mRNA splicing is entwined with the molecular details of ATPase/helicases function, which underlies the critical steps of spliceosome remodeling.

Briefly, allosterism is the phenomenon by which two distinct locations within a protein or protein complex become functionally coupled.[25] This coupling is accomplished by structural[26] and dynamic[27] changes that promotes the transfer of chemical information between the two physically separated regions. In allosterically-regulated enzymatic catalysis, binding of the effector at a distant site commands thermodynamic and kinetic specifications at the active site essential for the reaction.[28] Generally, these parameters make accessible the optimal conformation of the active site characteristic of the reactive state.[29,30] In the case of ATPases/helicases, the RNA strand binding is allosterically coupled to ATP hydrolysis. Through this allosteric communication between the RNA binding tunnel and the catalytic site, these enzymes carry out ATP hydrolysis, only upon RNA binding. This preserves the fidelity of function and avoids wasteful expenditure of ATP.[31]

Here, complementary classical and quantum/classical (QM/MM) Molecular Dynamics (MD) simulations enabled us to unprecedentedly unlock the molecular terms of Prp2 ATP hydrolysis and to pinpoint the allosteric mechanism underlying the RNA-mediated activation of its ATPse/helicase activity. The performed simulations and supplementary bioinformatic analyses captured the existence of key structural motifs underlying the RNA-driven activation, which are conserved across the DExH-box helicases family. The mechanism revealed herein extends the already known RNA-driven activation mechanisms of ATPase function, beyond the Glu-switch activation mechanism only operative for AAA+ enzymes.[32] Our study thrusts molecular understanding of this important family of RNA helicases involved in a wide spectrum of biomolecular processes. Since deregulated of RNA helicase activity is frequently linked to gene expression-associated diseases,[31] our study supplies a conceptual basis for future drug discovery campaigns aimed to develop RNA helicases inhibitors for the treatment of multiple disorders.

## 2. METHODS

**System Preparation.** The initial coordinates of the system were obtained from the crystallographic structure of Prp2 from *C. Thermophilum* in complex with RNA, $Mg^{2+}$ and an ATP analog (ADP-$BeF_3^-$, PDB code 6ZM2).[24] The ATP molecule was reconstructed by replacing the corresponding atoms in the crystallized analog. The coordinates of the missing atoms and the crystallographically unsolved hydrogens were added in sterically favorable positions. Protonation states for the ionizable residues were set up according to the results provided by the PROPKA3 server.[33] The reference pH value was 7 as adjusted experimentally in the kinetic studies.[15] Next, the system was solvated with a ~114 Å$^3$ cubic box of water molecules (TIP3P)[34] with a minimum distance of 15 Å between any protein atom and the edge of the box. The water molecules from the crystal structure were preserved. Finally, a total of 134 $Na^+$ ions and 121 $Cl^-$ ions were added to neutralize and to set an ion concentration of 150 mM. The number of ions was computed with the server and protocols provided by Schmit and colleagues.[35] The complete system contains ~130 k atoms. The topology was done using AmberTools17 tleap package.[36]

**Classical Molecular Dynamics (MD).** Protein, RNA and water molecules were modeled using AMBERff14SB,[36] OL3[37] and TIP3P[34] force-field parameters, respectively. Mono-valent ions were treated using the Joung-Cheatham[38] parameters while the $Mg^{2+}$ ion was treated using the Allnér-Nilsson-Villa parameters.[39] For ATP and ADP cofactors, parameters from Meagher-Redman-Carlson were employed.[40]

All MD simulations were performed in various steps with the following protocol: (i) A minimization is performed on the solvent molecules, mono-valent ions, and hydrogens using 2500 minimization steps with the conjugate gradient algorithm. (ii) A short 200 ps MD simulation of the solvent molecules and mono-valent ions is carried out, with the positions of the $Mg^{2+}$ ion and the positions of the heavy atoms of the protein, RNA and cofactors restrained. (iii) Two energy minimizations, one with the protein and RNA backbone restrained and another unrestrained are done. (iv) The whole system is heated gradually in three sequential NPT simulations from 100 K to 310 K, with a constant pressure of 1 bar. Finally, (v) four independent replicas of 1-µs long NPT MD simulations are performed at 310 K and 1 bar, by randomly reassigning velocities at the initial structure.

In the treatment of non-bonded interactions, the cutoff limit for short-range interactions was 10 Å whereas a Particle Mesh Ewald (PME) model was used for the long-range interactions.[41,42] Temperature control was performed using Langevin dynamics with a 0.5 $ps^{-1}$ collision frequency.[43,44] Pressure control was performed with Berendsen barostat with a relaxation time of 0.5 ps.[45] For all classical simulations, the SHAKE algorithm was used to constrain light-heavy atoms vibrations.[46,47] The Verlet algorithm was used to propagate the system.[48] All the classical calculations were run in the AMBER GPU software version and most of the analyses were performed using the CPPTRAJ package.[36,49,50]

To avoid artifacts induced by the inaccuracy of force-field based simulations, in all classical simulations a restraint was applied on the positions of the $Mg^{2+}$ ion and the six oxygen atoms coordinated to it. Different crystallographic structures[19,24,51] of Prp2 with ADP, ATP analogues, with and without RNA, were evaluated to verify that the coordination sphere remains invariant at different status of the enzyme. Hence, seven atoms ($Mg^{2+}$ ion, 3 oxygen atoms of the coordinated water molecules, the hydroxyl oxygen atom of Thr327 residue and two oxygen atoms of ATP) composing the $Mg^{2+}$ coordination sphere were restrained throughout the MD simulation.

Additionally, in the equilibration prior to the QM/MM MD simulations, restrains were applied on two distances that were considered chemically relevant and could not be modeled correctly by classical MD methods (i.e. O:W$^{cat}$-Pγ:ATP and H:W$^{cat}$-Oε:Gln621). All the above mentioned restrains were set up with a force constant of 350 kcal·mol$^{-1}$·Å$^{-2}$.

**QM/MM Molecular Dynamics.** Equilibrium QM/MM simulations on all the minima along the reaction mechanism were performed with CP2K v. 9.0 engine.[52] The same force field topology was used to describe the MM part whereas the gradient-corrected exchange-correlation Becke–Lee–Yang–Parr (BLYP) functional was employed for the QM region.[53,54] The DFT functional was supplemented with D3 dispersion corrections.[55] The Gaussian and plane waves approximation implemented in the Quickstep package was employed for the electronic structure calculations.[56] Double zeta plus polarization basis set[57] were used to represent the valence orbitals and the electron density was treated by plane waves with a 320 Ry cutoff. Gaussian-type pseudopotentials of Goedecker–Teter–Hutter were used to simulate the atomic cores.[58,59] Wavefunction optimization was performed by the orbital transformation approach with a convergence criterion of $1·10^{-6}$ for the electronic gradient.[60] The mechanical coupling was performed using the link atom procedure.[61,62] The electrostatic coupling was treated by the protocol introduced by Laino et al.[63] The simulations were performed in a NVT ensemble (310 K), employing a velocity rescaling thermostat. A time step of 0.5 fs was employed during the 10 ps long simulations. The QM region, comprising 95 atoms, is shown in Figure S1.

**QM/MM Metadynamics.** We performed QM/MM metadynamics simulations to compute the free energy profile along the ATP hydrolysis mechanism. This allows us to unlock the ATPase mechanism and see how this is activated by RNA binding. To this end, we carried out metadynamics simulations to explore the whole free energy profile in the presence of RNA. In addition, metadynamics simulations of the rate-determining step were also done in the absence of RNA. In the last case, the large flexibility of the general base makes any recrossing event extremely unlikely. For this reason, in the absence of the RNA strand, the free energy barrier of the rate-determining step was calculated by performing multiple metadynamics runs and by stopping the simulation after the transition state has been overcome. This approach has been successfully used in other computational studies.[64]

In the presence of the RNA, at each step of the catalytic mechanism, the height of the added hills was set to 0.6125 kcal·mol$^{-1}$ ($<kT$ at 310 K) and the width of the Gaussian functions was set up according with the oscillations of the collective variables (CVs) in unbiased QM/MM MD simulations.[65] We added Gaussians functions every 30 fs corresponding to 120 or 60 MD steps (depending on the time step used). For chemical steps we used a time step of 0.25 fs while for the proton networks rearrangements time step was set to 0.5 fs. As proposed by Ensing et. al., we stopped the QM/MM metadynamics after all the critical points (principal minima and transition states) were sampled at least once.[65] In each case, three replicas were carried out starting with half of the first minimum already filled. The height of the hills for those replicas was reduced to 3/4, 1/2 and 1/4 of the original value.

Since in the absence of RNA, the active site is more flexible, and the position of the general base is quite variable, we ran six replicas of the rate-determining step to assess the reproducibility of the results. A first complete filling was done using 0.6125 kcal·mol$^{-1}$ as the height of hill. Then, 5 extra replicas were done, starting with half of the first minimum already filled. The height of the hills for those replicas was 3/4, 1/2, 1/4, 1/6 and 1/8 of the original value.

In all cases, variances in the barriers, along the performed replicas using different hill heights, were used to calculate the error following the protocol proposed by Nair et al.[66] The molecular mechanism resulted the same for all the Gaussian heights used. Details of the CVs and parameters used in each metadynamic run can be found in the Supplementary Information Table S1. It is well known that metadynamics proceeds via the minimum free energy path. Therefore, by selecting proper collective variables, we can directly discriminate between different reaction pathways and disregard the less favorable ones.

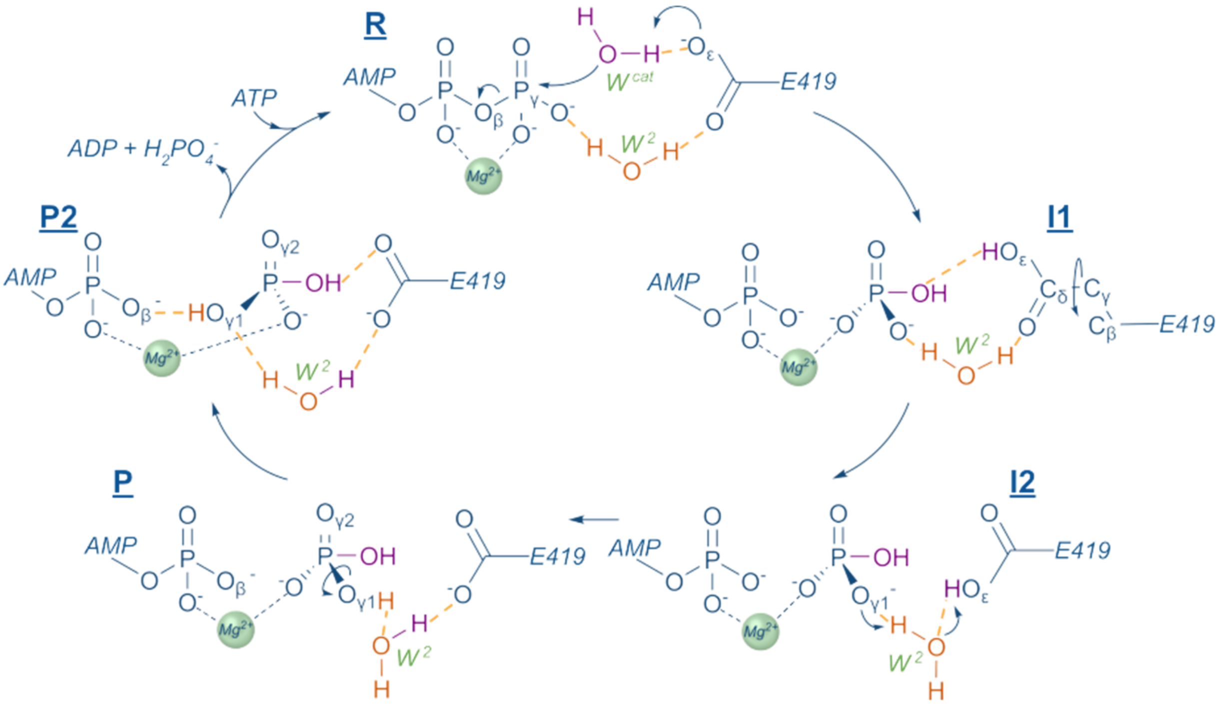
We remark that, differing with other studies,[67–69] here we did not consider an hydroxide ion as nucleophile as its diffusion from the bulk water to the active site and its positioning between an ionized Glu419 (general base) and the g-phosphate makes it extremely unlikely due to electrostatic repulsion.

## 3. RESULTS AND DISCUSSION

**Enzymatic mechanism of ATP hydrolysis in Prp2.** In order to elucidate the mechanism (Scheme 1) of ATP hydrolysis catalyzed by Prp2, we employed metadynamic (MTD) simulations at hybrid quantum/classic (QM/MM) theory level, based on DFT-BLYP-D3[53–55] for the QM part and on AMBER force fields for the MM part. The model system was built based on the crystallographic structure of Prp2 in complex with an ATP mimic (ADP-$BeF_3^-$, PDB ID: 6ZM2).[24] After an initial equilibration of 150 ns-long MD simulations, the system was further relaxed for 10 ps by a QM/MM MD simulation (Figure S2).

The equilibrated Michaelis complex (**R**, Figure 1) exhibited a stable structure similar to the crystallographic one (average RMSD of the active site = 1.65 Å) and maintained the same hydrogen (H)-bond network surrounding the ATP substrate. Namely, the Arg625 interacts with ATP γ-phosphate, while Lys326 and Arg628 interact with both the γ and β phosphates (Figure S1). Arg625, besides interacting with the γ phosphate of ATP, also interacts with the Oε:Gln621. The Glu419, Ser578 and Gln621, in turn, H-bond

**SCHEME 1. Sketch of the ATP hydrolysis reaction mechanism catalyzed by the DEAH-box Prp2 ATPase/helicase here proposed.**



with the catalytic water, hence contributing to retain the nucleophilic oxygen ($O_{nuc}$:Wcat) at the optimal distance (3.3±0.2 Å) from the Pγ:ATP and in line orientation (161±0.2°) for the nucleophilic attack. Consistently with the crystal structure, a second water molecule (W2) settles within the active site (Figure S1), acting as H-bond donor to the Oγ:ATP and the Oε:Glu419, while acting as H-bond acceptor of the H-Nε:Gln621 and the H-N:Ala451.

Next, we inspected the first step of the ATP hydrolysis reaction (**R**→**I1**, Figure 1 and Scheme 1) by promoting the nucleophilic attack of the $O_{nuc}$:Wcat on the Pγ:ATP in QM/MM MTD simulation. To this end, we used as collective variable (CV1, see Table S1) the difference between the distances of the forming bond (i.e. distance between $O_{nuc}$:Wcat and Pγ:ATP) and the breaking bond (i.e. the distance between Oβ:ATP and Pγ:ATP). The hydrolysis of phosphates proceeds through one of three possible mechanisms, dissociative, associative or concerted (see Scheme S1).[70,71] Here, we reveal that the hydrolysis of ATP occurs through a synchronous concerted path where the attack of the nucleophilic water on the scissile phosphate is in sync with the removal of one proton from Wcat by Glu419 residue, which acts as the general base of the catalytic process. Therefore, the use of CV1 allows us to discard competitive pathways that pass through the formation of intermediates, such as metaphosphates or pentacoordinated phosphates. Moreover, even in the absence of a CV specifically biasing the proton transfer, the deprotonation of the nucleophilic water was always mediated by Glu419 in the four replicas of the first reaction step we performed. Thus, the assistance of any other nearby bases is unlikely. The free energy surfaces and time evolution of the CVs for this and the rest of steps are shown in Figures S4 and S5, respectively.

First step, leading to the formation of a $HPO_4^{2-}$ and an ADP molecule, occurs by overcoming a Helmholtz free energy barrier ($\Delta A^{\ddagger}$) of 16.7±0.8 kcal·mol$^{-1}$ (Figure 1 and Scheme 1). Close to the transition state ($\mathbf{TS}^{R\rightarrow I1}$), the distances of the forming and breaking bonds ($O_{nuc}$:Wcat-Pγ:ATP and Pγ:ATP-Oβ:ATP) are both close to 2.2 Å (Figure S6). Notably, the calculated $\Delta A^{\ddagger}$ is in close agreement with the experimentally measured $k_{cat}$ of 3.2 s$^{-1}$,[21] which by applying the Eyring equation,[72] in the context of the Transition State Theory (TST),[73] corresponding to a Gibbs free energy barrier ($\Delta G^{\ddagger}$) of 17.4 kcal·mol$^{-1}$ at 310 K.

The catalytic reaction further proceeds (**I1**→**I2**) with a rearrangement of the H-bond network of the Glu419 carboxylic group. After undergoing a rotation of the Cβ-Cγ-Cδ-OHε:Glu419 dihedral angle by ~35º, HOε:Glu419 is oriented towards the oxygen of a second water molecule (W2, Figure 1 and Scheme 1) located in the active site. As a result, W2 also rearranges to H-bond to a negatively charged oxygen atom of the newly formed $HPO_4^{2-}$ ion. In this manner, W2 bridges the $HPO_4^{2-}$ ion and HOε:Glu419. We estimate that the Glu419 conformational rearrangement occurs at a $\Delta A^{\ddagger}$ of 4.6±0.7 kcal·mol$^{-1}$ ($\mathbf{TS}^{I1\rightarrow I2}$, Figures S4 and S5). This value was estimated by performing a MTD simulation, using as CV2 (Table S1) the difference of distances of the hydrogen of the general base HOε:Glu419 and the nucleophilic $O_{nuc}$:Wcat, which, after the first catalytic step, is part of the $HPO_4^{2-}$ ion, minus the distance of HOε:Glu419 and O:W2. As a result, a metastable intermediate (**I2**, Figure 1)

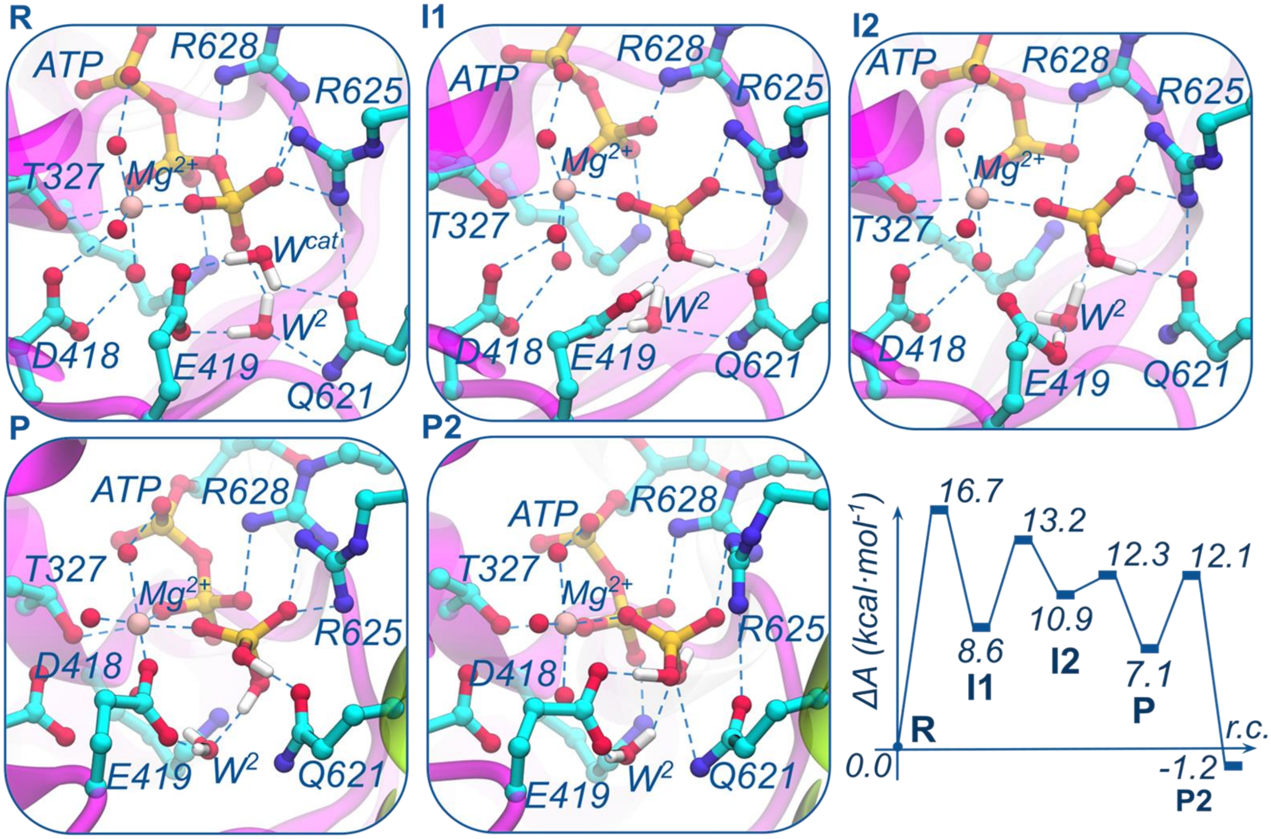


**FIGURE 1.** Representative structures of the key states of the ATP hydrolysis mechanism catalyzed by Prp2 and activated by the binding of an RNA strand. The Prp2 protein is shown in new cartoon with RecA1 and RecA2 domain colored in magenta and green, respectively. The active site residues are shown in ball and sticks and colored by atom name. Last panel: Schematic representation of the free energy profile of the ATP hydrolysis mechanism calculated at QM(BLYP-D3(DZVP))/MM level. Reaction free energies and free energy barriers were calculated with respect to the reactant state and are reported in kcal·mol$^{-1}$. A dynamic view of the mechanism is shown in Movie S1

is formed ($\Delta A$ = 2.3 kcal·mol$^{-1}$ from **I1**), which is rapidly converted into products (**P**, Figure 1 and Scheme 1) through a $\Delta A^{\ddagger}$ of 1.4 kcal·mol$^{-1}$ (**TS$^{I2\rightarrow P}$**).

The formation of products occurs via an asynchronous double proton transfer first from H:W2 to Oγ$_1$:HPO$_4^{2-}$ and then from HOε:Glu419 to O:W2. This process is achieved by performing a MTD simulation with two CVs (i.e. CV3 equal to the difference of the distances between HOε:Glu419 and Oε:Glu419 and the distance between HOε:Glu419 and O:W2, and CV3' equal to the difference of distances between H:W2 and O:W2 and the distance between H:W2 and Oγ$_1$:HPO$_4^{2-}$ (Figures S4 and S5 and table S1)). The almost spontaneous character of this process is due to the basicity of the HPO$_4^{2-}$ ion at physiological conditions. In addition to this pathway, we tried to evaluated the proton transfer directly to the other oxygen of HPO$_4^{2-}$ ion (the one coordinated to Mg$^{2+}$). However, because its participation in the coordination sphere decreases its basicity, this alternative path resulted in very high energies and distortion of the Mg$^{2+}$ ion coordination sphere. Furthermore, the crystal structures with ATP[24] and ADP[19,51] suggest that i) the second water molecule may participate in the ATP hydrolysis reaction and ii) the coordination sphere remains invariant after hydrolysis, supporting our findings.

**P** still lies at an $\Delta A$ of +7.1 kcal·mol$^{-1}$ as compared to reactant **R**. The formation of an energetically favorable product (**P2**, Figure 1 and Scheme 1), occurs only after an H-bond network rearrangement. Here the proton of the HOγ1:H$_2$PO$_4^-$ reorients towards the Oβ:ADP. A similar mechanism of product stabilizations was reported also for other ATPases, and may be instrumental to reduce the interaction of H$_2$PO$_4^-$ with the catalytic site residues, thus facilitating its release.[74,75] This rearrangement occurs by overcoming a $\Delta A^{\ddagger}$ of 5.0±1.3 kcal·mol$^{-1}$ and is achieved by performing a MTD simulation using as CV4 the dihedral angle HOγ$_1$-Oγ$_1$ -Pγ-Oγ$_2$:H$_2$PO$_4^-$, (Figures S4 and S5). This step leads to an overall exergonic catalytic process ($\Delta A$ = -1.2 kcal·mol$^{-1}$).

It is well known that Prp2 as well as all helicases of the DExH-box family are RNA-dependent ATPases.[21] In order to explore the role of the RNA strand in activating the Prp2

ATPase function we equilibrated the system with 1 μs-long MD simulation in the absence of the RNA and we performed 6 MTD replicas of the nucleophilic attack (i.e. the rate determining step). As a result, the calculated $\Delta A^{\ddagger}$ raises to 23.5±3.2 kcal·mol$^{-1}$ (Figure S4). The large standard deviation of this free energy barrier, by comparison with the value obtained in the presence of RNA, is due to the increased flexibility of the general base Glu419 (*vide infra*).

**The RNA client allosterically places the catalytic water molecule in a reactive configuration.** In order to inspect the way the single strand RNA allosterically triggers the activation of Prp2 ATPase/helicase function, we carried out four independent 1 μs-long classical MD simulations on the reactant state in the presence and absence of a seven nucleotides-long polyU RNA strand. Surprisingly, the structure of the active site is very similar in the two cases. Nonetheless, the RNA strand constraints the reactive disposition of Glu419 and Ser578 residues lining the active site. Indeed, the flexibility (root mean square fluctuation in Figure S3 and Table S1) of these residues markedly raises in the absence of the RNA strand. As detailed above, Glu419, besides taking direct part in most steps of the enzymatic mechanism, also properly orients the catalytic water Wcat for the nucleophilic attack (Figure 2 a, c), which is the rate determining step of the ATP hydrolysis. The increased flexibility of Glu419 and Ser578 remodels the hydrophilic cavity in the vicinity of Pγ:ATP, increasing its volume. This enables the entrance of an extra water molecule into the active site (W3, Figures 2b, d) and, most importantly, triggers a repositioning of the nucleophilic Wcat. Briefly, in the lack of the RNA strand, the Wcat no longer explores the optimal and indispensable in-line position for the nucleophilic attack (Figure 2).

In order to capture the molecular signature underlying this RNA-mediate allosteric activation of Prp2 catalysis, we closely inspected the residues mechanically coupling the catalytic site and the RNA binding tunnel. As a result, we observed that in the presence of the RNA strand Arg353 engages persistent H-bond and salt-bridge interactions (with its NH backbone and guanidinium moieties) to the phosphate backbone of U5 and U6 nucleotides, respectively. This in turn rigidifies the flanking Arg352 residue whose guanidinium moiety engages a salt-bridge to the side chain of Glu422 (Figure 3a, b). In Figure 3 is depicted the distance distribution between Cδ:Glu419-Cζ:Arg352 as D1 showing larger values in the absence of RNA (4.69±0.5 and 6.25±1.22 Å in the presence/absence of RNA, respectively, Figure 3a, b, e). Moreover, Glu422 and the Glu419 belong to the same loop and directly H-bonds through their backbone atoms. Then, the salt-bridge between Arg352 and Glu422 locks the position of the Glu419 in the presence of RNA, the general base of the ATPase reaction, into the optimal position to activate the nucleophilic water Wcat (Figure 3). While in the absence of RNA, Glu422 and Glu419 are more flexible and the latter does not establish a proper position to preorient and to activate the nucleophilic water Wcat to attack Pγ:ATP. This provides an explanation to the markedly larger free energy barrier calculated in the absence of the RNA (Figure S4).

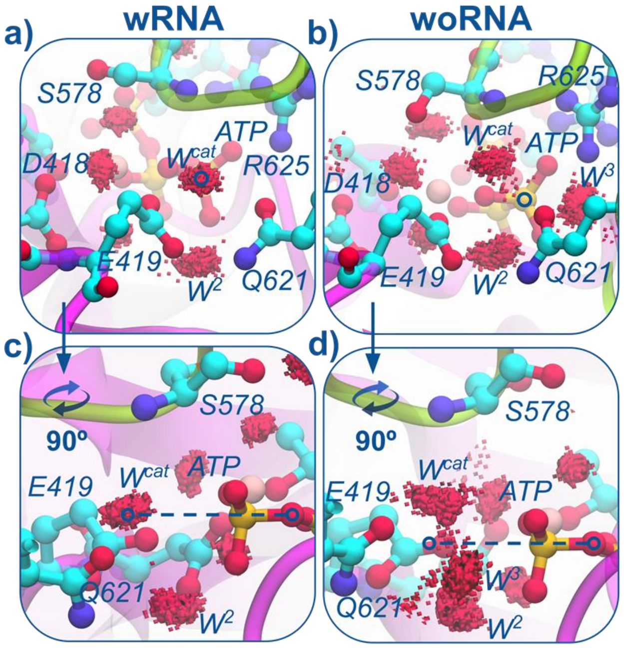


**FIGURE 2.** Representative frames of the orientation of key Prp2 residues and substrate. The Prp2 protein is shown in new cartoon with RecA1 and RecA2 domain colored in magenta and green, respectively. The active site residues are shown in ball and sticks and colored by atom name. Density distribution of the positions visited by the nucleophilic and accessory water molecules (Wcat, W2, W3) along the molecular dynamics simulations are show as red surface in presence (panels a and c) and absence (panels b and d) of a 7-nucleotides-long RNA strand. The ideal position of the nucleophilic water is depicted by a blue circle and the ideal alignment for the attack to the Pγ:ATP atom is highlighted by blue dashed lines. Density distributions for the replicas are reported in Figure S7.

Additionally, we identified a second set of important interactions. Namely, the side chain of Asn573 interacts with the sugar of U4 nucleotide, rigidifying the loop formed by residues 572 - 579 where Ser578 belongs to (Figure 3c, d). In the absence of the RNA strand HN:Ser578, which faces the active site, becomes more flexible (Figure S3 and Table S2) and lies at larger distances from Pγ:ATP (4.78±0.2 vs 5.07±0.2 Å in the presence/absence of RNA, respectively, D2 in Figure 3c, d, f). This shift in the relative position of Ser578 strongly contributes to the Wcat repositioning, thus leading to impaired catalysis. Remarkably, Asn573 and Arg353 interact with RNA through one sugar and two phosphates, respectively, which is congruent with the sequence-independent nature of Prp2.[18] The observed trends were confirmed in all MD simulations replicas and are reported in Figure S8 of the Supporting Information.

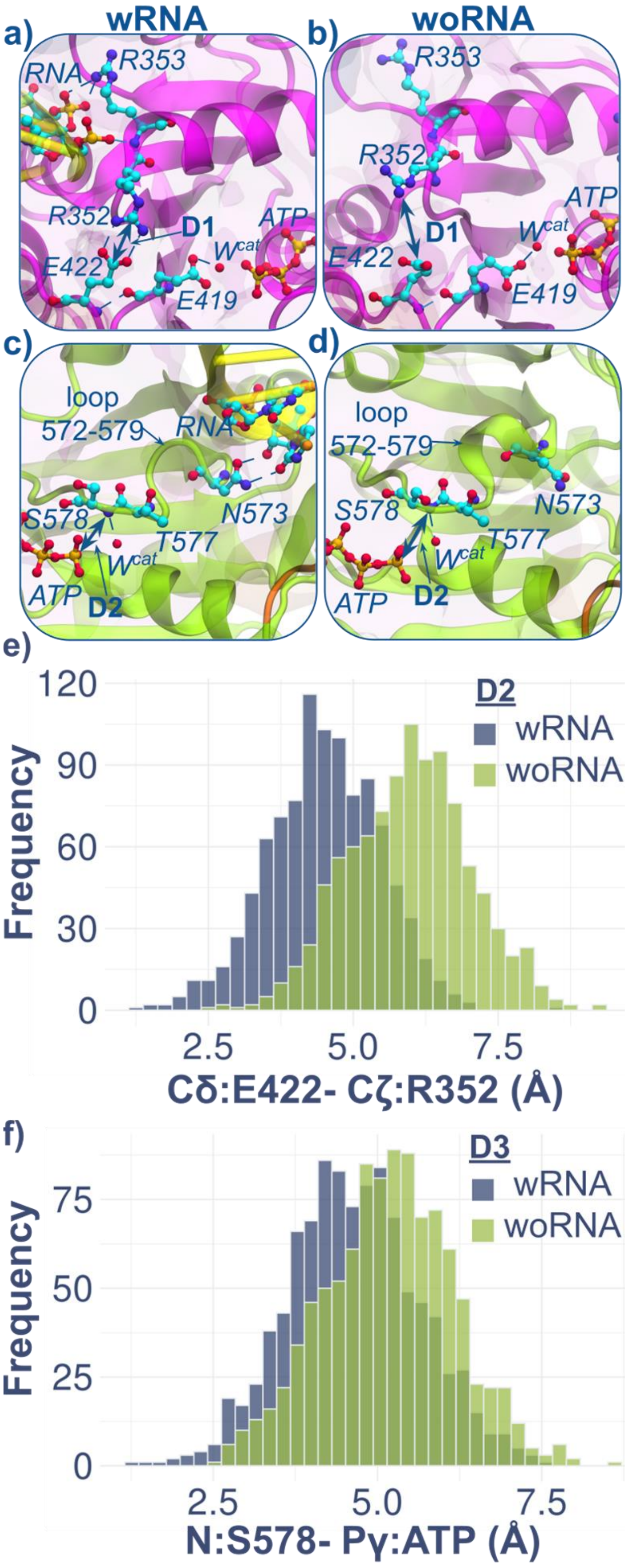


**FIGURE 3.** Representative frames with the orientation of the key Prp2 residues, ATP substrate and catalytic water in the presence (left) and absence (right) the RNA strand. The Prp2 protein is shown in new cartoon with RecA1 and RecA2 domain colored in magenta and green, respectively. The active site residues are shown in ball and sticks and colored by atom name. Orientation of the Arg352, Arg353, Glu422, Glu419 residues in a) presence and b) absence of the RNA strand. Orientation of Asn573 and the U4 base of RNA, Thr577 and the nucleophilic water in the c) presence and d) absence of the RNA strand. e) Distribution of the distance (Å) between Cδ:Glu419-Cζ:Arg352 atoms (marked as D1 in a) and b)) along the 1 μs-long classical MD trajectory in the presence (blue) and absence (green) of the RNA strand. f) Distribution of the distance (Å) between N:Ser578-Pγ:ATP atoms (marked as D2 in c) and d)) along the 1 μs-long classical MD in the presence (blue) and absence (green) of the RNA strand. A dynamic view of the interactions involved in the allosteric paths is shown in Movie S2.

A bioinformatic analysis done with the Protein BLAST server[76,77] strikingly revealed that the residues involved in the allosteric activation of Prp2 ATP hydrolysis catalysis are highly conserved among other DExH-box ATPase/helicases (Figure 4). Namely, the Arg352, Arg353, Glu419, Glu422 (hereafter name RR-EE motif), and the Asn573 and Ser578 (NS motif) residues are invariant among the functional orthologs of Prp2 and in the other spliceosomal DExH-box ATPase/helicases Prp16, Prp22, Prp43. The high degree of conservation suggests that any mutation (natural or *in vitro*) of the aminoacids involved either in allosteric communication from the RNA binding site to the catalytic site (Arg352, Arg353, Glu419, Glu422 and Asn573 and Ser578) or in the activation of the nucleophilic water (Glu419) directly impacts the ATPase/helicase activity of any member of the DExH-box family. Therefore, our study supplies a set of experimentally testable hypotheses to be performed in future studies aimed at decrypting the mechanism of DExH-box helicases.

The human genome encodes 16 DExH-box RNA helicases that are involved in a variety of key biological processes such as homologous recombination repair,[78] nonsense decay activation mediation,[79] unwinding of G4 structures forming in the 3′-untranslated regions of mRNA,[80] ribosome biogenesis, global translation, and mitochondrial metabolism.[81] Stunningly, the residues involved in allosteric activation of Prp2 ATPase function are conserved even among several other DExH-box ATPase/helicases not taking part to the splicing cycle (Figure 4), suggesting that this allosteric regulation may be broadly applicable to all DExH helicases. This mechanism is nonetheless unique to the DExH subfamily as revealed by the sequence alignment of other spliceosomal helicases from the DExD-box and Ski2-like families (Figure S7), where instead these key motifs are lacking. More broadly, this mechanism resembles the so called ‘glutamate switch’ underlying the allosteric activation in AAA+ family of proteins.[32] In AAA+ enzymes, a positively charged Arg/Lys or polar Gln/Asn/Thr/Ser act as Glu-switch residues are proposed to thwart ATP hydrolysis by H-bonding to a catalytic Glu residues and holding it in a catalytically inactive conformation until the external

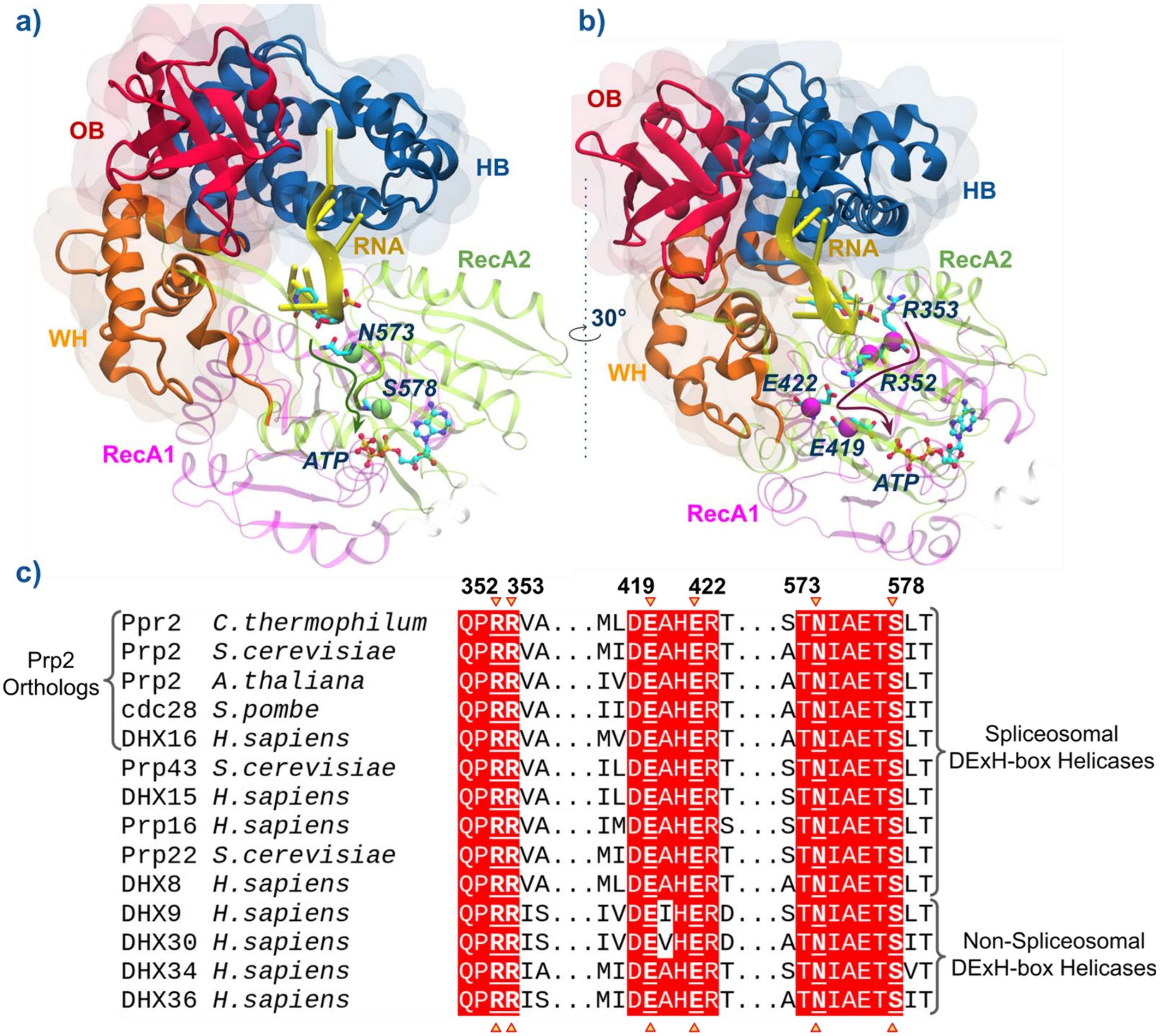


**FIGURE 4.** Summary of the key interaction tracing the activation path from key RNA stabilizing residues to the catalytic moiety. The Prp2 protein is shown in new cartoon with RecA1, RecA2 WH, OB and HB domains are colored in magenta, green, orange, red and blue, respectively. The residues involved in allosteric activation of Prp2 ATPase/helicase function are shown in licorice colored by atom name with their Cα highlighted with a van deer Waals (vdw) sphere. a) Front view showing the path connecting Asn573 to Ser578 b) side view showing the path connecting Arg353 to the catalytic residue Glu419. The Cα of key residues involved in the hydrogen bond networks connecting the RNA binding tunnel to the catalytic site are shown in green and magenta vdw spheres (panels a) and b), respectively). c) Sequence alignment among Prp2 and various DExH-box ATPases/helicases. The conserved residues are shaded red and those involved in allosteric modulation are marked in bold and underlined.

signal (nucleic acid and substrate ATP binding) occurs. In DExH-box helicases, instead, RNA binding/dissociation switches on/off the salt-bridge between Arg and Glu residues, in the center of a 'RR-EE' motif where the first residue directly interacts with RNA and the latter is the catalytic Glu419 that acts as base. The formation of this salt-bridges contributes to mechanically regulating the RNA-driven activation of this pivotal protein family.

## 4. CONCLUSIONS

Here, we unprecedentedly disclose the mechanism of RNA-driven ATP hydrolysis of Prp2. By performing QM/MM metadynamics simulations, we unlock that the ATP hydrolysis mechanism proceeds in four steps with the rate determining one being the attack of the nucleophilic water to the Pγ:ATP ($\Delta A^{\ddagger}$ = 16.7±0.8 kcal·mol$^{-1}$, in agreement with experimental kinetic report[21] of $k_{cat}$ 3.2 s$^{-1}$, 17.4 kcal·mol$^{-1}$ at 310 K). This is followed by a series of proton transfer and H-bond rearrangements, which lead to the exergonic formation of the $H_2PO_4^-$/ADP products ($\Delta A = -1.2$ kcal·mol$^{-1}$). Remarkably, we also reveal that in the absence of the RNA strand a marked increase in the activation free energy barrier of the rate determining step occurs ($\Delta A^\ddagger = 23.5 \pm 3.2$ kcal·mol$^{-1}$), thus confirming the role of the RNA in optimal positioning the nucleophilic water in a reactive conformation in the active site. This mechanistic hypothesis agrees with the available experimental evidence, both kinetic[21] and crystallographic[19,24,51].

Complementary, MD simulations disclosed the molecular terms of RNA-driven activation of ATP hydrolysis, elucidating that Arg353, by engaging strong interactions with the phosphate groups of the RNA, induces the formation of a stable salt-bridge interaction between Arg352 backbone and Glu422. When this interaction is formed, the flanking Glu419, the general base of the enzymatic process, is placed in an optimal position to activate the nucleophilic water. Concomitantly, the interactions between the Asn573 residue and the RNA sugar of the U3 base affects the volume of the catalytic site, also contributing to optimally placing the nucleophile for the reaction. In the absence of the RNA strand, these key interactions are lost, triggering an increased flexibility and increasing the volume of the active site. As a result, the catalytic pocket experiences a displacement of nucleophilic water and the entry of an additional water molecule hindering catalysis.

In RNA ATPase/helicase activity is often tightly coupled to interactions with client substrate, but the mechanisms behind this linkage are frequently unclear. A bioinformatic analysis reveals that this RNA-driven ATPase activation is conserved only across the DExH-box RNA helicases which, besides participating in splicesome remodeling, take part in other key cellular processes.[31] The allosteric activation mechanism discovered here expands the Glu switch mechanism operative for AAA+ enzymes.[32] We suggested that in DExH-box helicases, an expanded RR-EE switch motif plastically regulates the ATPase/helicase function in response to the binding of an effector RNA strand. The high degree of conservation within the DExH-box family makes it possible to suggest that mutations over the RR-EE and NS motifs would negatively impact the proper function of these enzymes.

Our study advances the molecular understanding of the function of this key family of RNA helicases implicated across a wide range of vital cellular processes. Dysregulation of RNA helicase activity is often associated with tumorigenesis and genetic diseases.[31] In this scenario, our study provides a conceptual basis for future drug discovery efforts aimed at developing and probing the therapeutic relevance of RNA helicase inhibitors in the treatment of various disorders.

## ACKNOWLEDGEMENTS

We would like to thank the Spanish Ministerio de Ciencia e Innovación (grant PGC2021-23332OB-C21), the Generalitat Valenciana (PROMETEO, grant ref. CIPROM/2021/079) and Universitat Jaume I (grants UJI-2020-03 and UJI-2019-43). SM thanks the Generalitat Valenciana for a Grisolia Ph.D. grant (GRISOLIAP/2019/064) and HPC Europa 3 (HPC17RZ6NO) for the grant of mobility. The authors acknowledge to CINECA, for the availability of high performance computing resources and support as well as the Servei d'Informàtica of Universitat Jaume I. AM thanks the Italian Association for Cancer Research (AIRC, IG grant 24514) for financial support.

## ASSOCIATED CONTENT

## SUPPORTING INFORMATION

- Snapshot of the active site of Prp2 and explicit representation of the QM region set up for the QM/MM simulations; root mean square deviation (RMSD) analyses for classical and hybrid MD simulations; root mean square fluctuation (RMSF) analyses for classical MD simulations; free energy profiles of the evaluated reaction mechanism steps; time evolution of the collective variables (CVs) along the metadynamics trajectories; time evolution of main distances along the metadynamics simulation of the rate determinant step; water density analysis at the active site and distance distribution of RR-EE and NS motifs for the performed replicas; sequence alignment among Prp2, various DExH-box ATPases/helicases, spliceosomal DExD-box ATPases/helicases and Brr2 helicase of the Ski-like family; detailed description of the collective variables and parameters of the metadynamics simulations; More-O'Ferrall–Jencks representation of the lysis pathways for ATP hydrolysis. (PDF)
- Movies showing the ATP hydrolysis mechanism and the allosteric RR-EE/NS motifs underlying RNA-driven activation of Prp2 catalysis (MP4)
- Representative structure of the precatalytic reactive complex of Prp2 with ATP and RNA. (PDB)

## AUTHOR INFORMATION

### Corresponding Author

**Alessandra Magistrato** – [2]Department National Research Council of Italy (CNR) - Institute of Material (IOM) c/o International School for Advanced Studies (SISSA), Trieste, Italy; Email: alessandra.magistrato@sissa.it

### Authors

**Santiago Movilla** – BioComp Group, Institute of Advanced Materials (INAM), Universidad Jaume I, 12071 Castellón, Spain.

**Maite Roca** – BioComp Group, Institute of Advanced Materials (INAM), Universidad Jaume I, 12071 Castellón, Spain; orcid.org/0000-0003-0937-4722

**Vicent Moliner** – BioComp Group, Institute of Advanced Materials (INAM), Universidad Jaume I, 12071 Castellón, Spain; orcid.org/0000-0002-3665-3391

**Notes**

The authors declare no competing financial interest.

## 5. REFERENCES

(1) Brody, E.; Abelson, J. The "Spliceosome": Yeast Pre-Messenger RNA Associates with a 40 S Complex in a Splicing-Dependent Reaction. *Science.* **1985**, *228* (4702), 963–967. https://doi.org/10.1126/science.3890181.

(2) Frendewey, D.; Keller, W. Stepwise Assembly of a Pre-MRNA Splicing Complex Requires U-SnRNPs and Specific Intron Sequences. *Cell* **1985**, *42* (1), 355–367. https://doi.org/10.1016/S0092-8674(85)80131-8.

(3) Grabowski, P.; Seiler, S.; Sharp, P. A Multicomponent Complex Is Involved in the Splicing of Messenger RNA Precursors. *Cell* **1985**, *42* (1), 345–353. https://doi.org/10.1016/S0092-8674(85)80130-6.

(4) Wahl, M. C.; Will, C. L.; Lührmann, R. The Spliceosome: Design Principles of a Dynamic RNP Machine. *Cell* **2009**, *136* (4), 701–718. https://doi.org/10.1016/j.cell.2009.02.009.

(5) Shi, Y. The Spliceosome: A Protein-Directed Metalloribozyme. *J. Mol. Biol.* **2017**, *429* (17), 2640–2653. https://doi.org/10.1016/j.jmb.2017.07.010.

(6) Staley, J. P.; Guthrie, C. Mechanical Devices of the Spliceosome: Motors, Clocks, Springs, and Things. *Cell* **1998**, *92* (3), 315–326. https://doi.org/10.1016/S0092-8674(00)80925-3.

(7) Borišek, J.; Casalino, L.; Saltalamacchia, A.; Mays, S. G.; Malcovati, L.; Magistrato, A. Atomic-Level Mechanism of Pre-MRNA Splicing in Health and Disease. *Acc. Chem. Res.* **2021**, *54* (1), 144–154. https://doi.org/10.1021/acs.accounts.0c00578.

(8) Casalino, L.; Palermo, G.; Rothlisberger, U.; Magistrato, A. Who Activates the Nucleophile in Ribozyme Catalysis? An Answer from the Splicing Mechanism of Group II Introns. *J. Am. Chem. Soc.* **2016**, *138* (33), 10374–10377. https://doi.org/10.1021/jacs.6b01363.

(9) Borišek, J.; Magistrato, A. All-Atom Simulations Decrypt the Molecular Terms of RNA Catalysis in the Exon-Ligation Step of the Spliceosome. *ACS Catal.* **2020**, *10* (9), 5328–5334. https://doi.org/10.1021/acscatal.0c00390.

(10) Borišek, J.; Magistrato, A. An Expanded Two-Zn$^{2+}$ -Ion Motif Orchestrates Pre-MRNA Maturation in the 3′-End Processing Endonuclease Machinery. *ACS Catal.* **2021**, *11* (7), 4319–4326. https://doi.org/10.1021/acscatal.0c05594.

(11) Will, C. L.; Luhrmann, R. Spliceosome Structure and Function. *Cold Spring Harb. Perspect. Biol.* **2011**, *3* (7), a003707–a003707. https://doi.org/10.1101/cshperspect.a003707.

(12) Cordin, O.; Beggs, J. D. RNA Helicases in Splicing. *RNA Biol.* **2013**, *10* (1), 83–95. https://doi.org/10.4161/rna.22547.

(13) He, Y.; Andersen, G. R.; Nielsen, K. H. Structural Basis for the Function of DEAH Helicases. *EMBO Rep.* **2010**, *11* (3), 180–186. https://doi.org/10.1038/embor.2010.11.

(14) Walbott, H.; Mouffok, S.; Capeyrou, R.; Lebaron, S.; Humbert, O.; van Tilbeurgh, H.; Henry, Y.; Leulliot, N. Prp43p Contains a Processive Helicase Structural Architecture with a Specific Regulatory Domain. *EMBO J.* **2010**, *29* (13), 2194–2204. https://doi.org/10.1038/emboj.2010.102.

(15) Kastner, B.; Will, C. L.; Stark, H.; Lührmann, R. Structural Insights into Nuclear Pre-MRNA Splicing in Higher Eukaryotes. *Cold Spring Harb. Perspect. Biol.* **2019**, *11* (11), a032417. https://doi.org/10.1101/cshperspect.a032417.

(16) Plaschka, C.; Newman, A. J.; Nagai, K. Structural Basis of Nuclear Pre-MRNA Splicing: Lessons from Yeast. *Cold Spring Harb. Perspect. Biol.* **2019**, *11* (5), a032391. https://doi.org/10.1101/cshperspect.a032391.

(17) Combs, D. J.; Nagel, R. J.; Ares, M.; Stevens, S. W. Prp43p Is a DEAH-Box Spliceosome Disassembly Factor Essential for Ribosome Biogenesis. *Mol. Cell. Biol.* **2006**, *26* (2), 523–534. https://doi.org/10.1128/MCB.26.2.523-534.2006.

(18) De Bortoli, F.; Espinosa, S.; Zhao, R. DEAH-Box RNA Helicases in Pre-MRNA Splicing. *Trends Biochem. Sci.* **2021**, *46* (3), 225–238. https://doi.org/10.1016/j.tibs.2020.10.006.

(19) Bai, R.; Wan, R.; Yan, C.; Jia, Q.; Lei, J.; Shi, Y. Mechanism of Spliceosome Remodeling by the ATPase/Helicase Prp2 and Its Coactivator Spp2. *Science.* **2021**, *371* (6525). https://doi.org/10.1126/science.abe8863.

(20) Roy, J.; Kim, K.; Maddock, J. R.; Anthony, J. G.; Woolford, J. L. The Final Stages of Spliceosome Maturation Require Spp2p That Can Interact with the DEAH Box Protein Prp2p and Promote Step 1 of Splicing. *RNA* **1995**, *1* (4), 375–390.

(21) Warkocki, Z.; Schneider, C.; Mozaffari-Jovin, S.; Schmitzová, J.; Höbartner, C.; Fabrizio, P.; Lührmann, R. The G-Patch Protein Spp2 Couples the Spliceosome-Stimulated ATPase Activity of the DEAH-Box Protein Prp2 to Catalytic Activation of the Spliceosome. *Genes Dev.* **2015**, *29* (1), 94–107. https://doi.org/10.1101/gad.253070.114.

(22) Warkocki, Z.; Odenwälder, P.; Schmitzová, J.; Platzmann, F.; Stark, H.; Urlaub, H.; Ficner, R.; Fabrizio, P.; Lührmann, R. Reconstitution of Both Steps of Saccharomyces Cerevisiae Splicing with Purified Spliceosomal Components. *Nat. Struct. Mol. Biol.* **2009**, *16* (12), 1237–1243. https://doi.org/10.1038/nsmb.1729.

(23) Kim, S. H.; Lin, R. J. Spliceosome Activation by PRP2 ATPase Prior to the First Transesterification Reaction of Pre-MRNA Splicing. *Mol. Cell. Biol.* **1996**, *16* (12), 6810–6819. https://doi.org/10.1128/MCB.16.12.6810.

(24) Hamann, F.; Zimmerningkat, L. C.; Becker, R. A.; Garbers, T. B.; Neumann, P.; Hub, J. S.; Ficner, R. The Structure of Prp2 Bound to RNA and ADP-$BeF_3^-$ Reveals Structural Features Important for RNA Unwinding by DEAH-Box ATPases. *Acta Crystallogr. Sect. D Struct. Biol.* **2021**, *77* (4), 496–509. https://doi.org/10.1107/S2059798321001194.

(25) Tsai, C.-J.; Nussinov, R. A Unified View of "How Allostery Works." *PLoS Comput. Biol.* **2014**, *10* (2), e1003394. https://doi.org/10.1371/journal.pcbi.1003394.

(26) Monod, J.; Wyman, J.; Changeux, J.-P. On the Nature of Allosteric Transitions: A Plausible Model. *J. Mol. Biol.* **1965**, *12* (1), 88–118. https://doi.org/10.1016/S0022-2836(65)80285-6.

(27) Cooper, A.; Dryden, D. T. F. Allostery without Conformational Change. *Eur. Biophys. J.* **1984**, *11* (2), 103–109. https://doi.org/10.1007/BF00276625.

(28) Lisi, G. P.; Loria, J. P. Allostery in Enzyme Catalysis. *Curr. Opin. Struct. Biol.* **2017**, *47*, 123–130. https://doi.org/10.1016/j.sbi.2017.08.002.

(29) Nussinov, R.; Tsai, C.-J. Allostery without a Conformational Change? Revisiting the Paradigm. *Curr. Opin. Struct. Biol.* **2015**, *30*, 17–24. https://doi.org/10.1016/j.sbi.2014.11.005.

(30) Jiménez-Osés, G.; Osuna, S.; Gao, X.; Sawaya, M. R.; Gilson, L.; Collier, S. J.; Huisman, G. W.; Yeates, T. O.; Tang, Y.; Houk, K. N. The Role of Distant Mutations and Allosteric Regulation on LovD Active Site Dynamics. *Nat. Chem. Biol.* **2014**, *10* (6), 431–436. https://doi.org/10.1038/nchembio.1503.

(31) Andrisani, O.; Liu, Q.; Kehn, P.; Leitner, W. W.; Moon, K.; Vazquez-Maldonado, N.; Fingerman, I.; Gale, M. Biological Functions of DEAD/DEAH-Box RNA Helicases in Health and Disease. *Nat. Immunol.* **2022**, *23* (3), 354–357. https://doi.org/10.1038/s41590-022-01149-7.

(32) Zhang, X.; Wigley, D. B. The "glutamate Switch" Provides a Link between ATPase Activity and Ligand Binding in AAA+ Proteins. *Nat. Struct. Mol. Biol.* **2008**, *15* (11), 1223–1227. https://doi.org/10.1038/nsmb.1501.

(33) Olsson, M. H. M.; Søndergaard, C. R.; Rostkowski, M.; Jensen, J. H. PROPKA3: Consistent Treatment of Internal and Surface Residues in Empirical p K a Predictions. *J. Chem. Theory Comput.* **2011**, *7* (2), 525–537. https://doi.org/10.1021/ct100578z.

(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.* **1983**, *79* (2), 926–935. https://doi.org/10.1063/1.445869.

(35) Schmit, J. D.; Kariyawasam, N. L.; Needham, V.; Smith, P. E. SLTCAP: A Simple Method for Calculating the Number of Ions Needed for MD Simulation. *J. Chem. Theory Comput.* **2018**, *14* (4), 1823–1827. https://doi.org/10.1021/acs.jctc.7b01254.

(36) Salomon-Ferrer, R.; Case, D. A.; Walker, R. C. An Overview of the Amber Biomolecular Simulation Package. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2013**, *3* (2), 198–210. https://doi.org/10.1002/wcms.1121.

(37) Pérez, A.; Marchán, I.; Svozil, D.; Sponer, J.; Cheatham, T. E.; Laughton, C. A.; Orozco, M. Refinement of the AMBER Force Field for Nucleic Acids: Improving the Description of α/γ Conformers. *Biophys. J.* **2007**, *92* (11), 3817–3829. https://doi.org/10.1529/biophysj.106.097782.

(38) Joung, I. S.; Cheatham, T. E. Determination of Alkali and Halide Monovalent Ion Parameters for Use in Explicitly Solvated Biomolecular Simulations. *J. Phys. Chem. B* **2008**, *112* (30), 9020–9041. https://doi.org/10.1021/jp8001614.

(39) Allnér, O.; Nilsson, L.; Villa, A. Magnesium Ion–Water Coordination and Exchange in Biomolecular Simulations. *J. Chem. Theory Comput.* **2012**, *8* (4), 1493–1502. https://doi.org/10.1021/ct3000734.

(40) Meagher, K. L.; Redman, L. T.; Carlson, H. A. Development of Polyphosphate Parameters for Use with the AMBER Force Field. *J. Comput. Chem.* **2003**, *24* (9), 1016–1025. https://doi.org/10.1002/jcc.10262.

(41) Essmann, U.; Perera, L.; Berkowitz, M. L.; Darden, T.; Lee, H.; Pedersen, L. G. A Smooth Particle Mesh Ewald Method. *J. Chem. Phys.* **1995**, *103* (19), 8577–8593. https://doi.org/10.1063/1.470117.

(42) Darden, T.; York, D.; Pedersen, L. Particle Mesh Ewald: An N ·log( N ) Method for Ewald Sums in Large Systems. *J. Chem. Phys.* **1993**, *98* (12), 10089–10092. https://doi.org/10.1063/1.464397.

(43) Uberuaga, B. P.; Anghel, M.; Voter, A. F. Synchronization of Trajectories in Canonical Molecular-Dynamics Simulations: Observation, Explanation, and Exploitation. *J. Chem. Phys.* **2004**, *120* (14), 6363–6374. https://doi.org/10.1063/1.1667473.

(44) Sindhikara, D. J.; Kim, S.; Voter, A. F.; Roitberg, A. E. Bad Seeds Sprout Perilous Dynamics: Stochastic Thermostat Induced Trajectory Synchronization in Biomolecules. *J. Chem. Theory Comput.* **2009**, *5* (6), 1624–1631. https://doi.org/10.1021/ct800573m.

(45) Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; DiNola, A.; Haak, J. R. Molecular Dynamics with Coupling to an External Bath. *J. Chem. Phys.* **1984**, *81* (8), 3684–3690. https://doi.org/10.1063/1.448118.

(46) Ryckaert, J.-P.; Ciccotti, G.; Berendsen, H. J. . Numerical Integration of the Cartesian Equations of Motion of a System with Constraints: Molecular Dynamics of n-Alkanes. *J. Comput. Phys.* **1977**, *23* (3), 327–341. https://doi.org/10.1016/0021-9991(77)90098-5.

(47) Kräutler, V.; van Gunsteren, W. F.; Hünenberger, P. H. A Fast SHAKE Algorithm to Solve Distance Constraint Equations for Small Molecules in Molecular Dynamics Simulations. *J. Comput. Chem.* **2001**, *22* (5), 501–508. https://doi.org/10.1002/1096-987X(20010415)22:5<501::AID-JCC1021>3.0.CO;2-V.

(48) Verlet, L. Computer "Experiments" on Classical Fluids. I. Thermodynamical Properties of Lennard-Jones Molecules. *Phys. Rev.* **1967**, *159* (1), 98–103. https://doi.org/10.1103/PhysRev.159.98.

(49) 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.* **2013**, *9* (9), 3878–3888. https://doi.org/10.1021/ct400314y.

(50) Roe, D. R.; Cheatham, T. E. PTRAJ and CPPTRAJ: Software for Processing and Analysis of Molecular Dynamics Trajectory Data. *J. Chem. Theory Comput.* **2013**, *9* (7), 3084–3095. https://doi.org/10.1021/ct400341p.

(51) Schmitt, A.; Hamann, F.; Neumann, P.; Ficner, R. Crystal Structure of the Spliceosomal DEAH-Box ATPase Prp2. *Acta Crystallogr. Sect. D Struct. Biol.* **2018**, *74* (7), 643–654. https://doi.org/10.1107/S2059798318006356.

(52) Hutter, J.; Iannuzzi, M.; Schiffmann, F.; VandeVondele, J. cp2k: Atomistic Simulations of Condensed Matter Systems. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2014**, *4* (1), 15–25. https://doi.org/10.1002/wcms.1159.

(53) 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* **1988**, *37* (2), 785–789. https://doi.org/10.1103/PhysRevB.37.785.

(54) Becke, A. D. Density-Functional Exchange-Energy Approximation with Correct Asymptotic Behavior. *Phys. Rev. A* **1988**, *38* (6), 3098–3100. https://doi.org/10.1103/PhysRevA.38.3098.

(55) Grimme, S.; Ehrlich, S.; Goerigk, L. Effect of the Damping Function in Dispersion Corrected Density Functional Theory. *J. Comput. Chem.* **2011**, *32* (7), 1456–1465. https://doi.org/10.1002/jcc.21759.

(56) VandeVondele, J.; Krack, M.; Mohamed, F.; Parrinello, M.; Chassaing, T.; Hutter, J. QUICKSTEP: Fast and Accurate Density Functional Calculations Using a Mixed Gaussian and Plane Waves Approach. *Comput. Phys. Commun.* **2005**, *167* (2), 103–128. https://doi.org/10.1016/j.cpc.2004.12.014.

(57) VandeVondele, J.; Hutter, J. Gaussian Basis Sets for Accurate Calculations on Molecular Systems in Gas and Condensed Phases. *J. Chem. Phys.* **2007**, *127* (11), 114105. https://doi.org/10.1063/1.2770708.

(58) Krack, M. Pseudopotentials for H to Kr Optimized for Gradient-Corrected Exchange-Correlation Functionals. *Theor. Chem. Acc.* **2005**, *114* (1–3), 145–152. https://doi.org/10.1007/s00214-005-0655-y.

(59) Hartwigsen, C.; Goedecker, S.; Hutter, J. Relativistic Separable Dual-Space Gaussian Pseudopotentials from H to Rn. *Phys. Rev. B - Condens. Matter Mater. Phys.* **1998**, *58* (7), 3641–3662. https://doi.org/10.1103/PhysRevB.58.3641.

(60) VandeVondele, J.; Hutter, J. An Efficient Orbital Transformation Method for Electronic Structure Calculations. *J. Chem. Phys.* **2003**, *118* (10), 4365–4369. https://doi.org/10.1063/1.1543154.

(61) Field, M. J.; Bash, P. A.; Karplus, M. A Combined Quantum Mechanical and Molecular Mechanical Potential for Molecular Dynamics Simulations. *J. Comput. Chem.* **1990**, *11* (6), 700–733. https://doi.org/10.1002/jcc.540110605.

(62) Davis, T. D.; Maggiora, G. M.; Christoffersen, R. E. Ab Initio Calculations on Large Molecules Using Molecular Fragments. Unrestricted Hartree-Fock Calculations of the Low-Lying States of Formaldehyde and Its Radical Ions. *J. Am. Chem. Soc.* **1974**, *96* (26), 7878–7887. https://doi.org/10.1021/ja00833a007.

(63) Laino, T.; Mohamed, F.; Laio, A.; Parrinello, M. An Efficient Real Space Multigrid QM/MM Electrostatic Coupling. *J. Chem. Theory Comput.* **2005**, *1* (6), 1176–1184. https://doi.org/10.1021/ct050123f.

(64) Bisha, I.; Laio, A.; Magistrato, A.; Giorgetti, A.; Sgrignani, J. A Candidate Ion-Retaining State in the Inward-Facing Conformation of Sodium/Galactose Symporter: Clues from Atomistic Simulations. *J. Chem. Theory Comput.* **2013**, *9* (2), 1240–1246. https://doi.org/10.1021/ct3008233.

(65) Ensing, B.; Laio, A.; Parrinello, M.; Klein, M. L. A Recipe for the Computation of the Free Energy Barrier and the Lowest Free Energy Path of Concerted Reactions. *J. Phys. Chem. B* **2005**, *109* (14), 6676–6687. https://doi.org/10.1021/jp045571i.

(66) Nair, N. N.; Schreiner, E.; Marx, D. Peptide Synthesis in Aqueous Environments: The Role of Extreme Conditions on Amino Acid Activation. *J. Am. Chem. Soc.* **2008**, *130* (43), 14148–14160. https://doi.org/10.1021/ja802370c.

(67) Ribeiro, A. J. M.; Ramos, M. J.; Fernandes, P. A. The Catalytic Mechanism of HIV-1 Integrase for DNA 3′-End Processing Established by QM/MM Calculations. *J. Am. Chem. Soc.* **2012**, *134* (32), 13436–13447. https://doi.org/10.1021/ja304601k.

(68) Wilson, K. A.; Fernandes, P. A.; Ramos, M. J.; Wetmore, S. D. Exploring the Identity of the General Base for a DNA Polymerase Catalyzed Reaction Using QM/MM: The Case Study of Human Translesion Synthesis Polymerase η. *ACS Catal.* **2019**, *9* (3), 2543–2551. https://doi.org/10.1021/acscatal.8b04889.

(69) Sgrignani, J.; Magistrato, A. QM/MM MD Simulations on the Enzymatic Pathway of the Human Flap Endonuclease (HFEN1) Elucidating Common Cleavage Pathways to RNase H Enzymes. *ACS Catal.* **2015**, *5* (6), 3864–3875. https://doi.org/10.1021/acscatal.5b00178.

(70) López-Canut, V.; Martí, S.; Bertrán, J.; Moliner, V.; Tuñón, I. Theoretical Modeling of the Reaction Mechanism of Phosphate Monoester Hydrolysis in Alkaline Phosphatase. *J. Phys. Chem. B* **2009**, *113* (22), 7816–7824. https://doi.org/10.1021/jp901444g.

(71) Zalatan, J. G.; Fenn, T. D.; Brunger, A. T.; Herschlag, D. Structural and Functional Comparisons of Nucleotide Pyrophosphatase/Phosphodiesterase and Alkaline Phosphatase: Implications for Mechanism and Evolution. *Biochemistry* **2006**, *45* (32), 9788–9803. https://doi.org/10.1021/bi060847t.

(72) Eyring, H. The Activated Complex in Chemical Reactions. *J. Chem. Phys.* **1935**, *3* (2), 107–115. https://doi.org/10.1063/1.1749604.

(73) Truhlar, D. G.; Hase, W. L.; Hynes, J. T. Current Status of Transition-State Theory. *J. Phys. Chem.* **1983**, *87* (26), 5523–5523. https://doi.org/10.1021/j150644a044.

(74) Li, G.; Cui, Q. Mechanochemical Coupling in Myosin: A Theoretical Analysis with Molecular Dynamics and Combined QM/MM Reaction Path Calculations. *J. Phys. Chem. B* **2004**, *108* (10), 3342–3357. https://doi.org/10.1021/jp0371783.

(75) Hayashi, S.; Ueno, H.; Shaikh, A. R.; Umemura, M.; Kamiya, M.; Ito, Y.; Ikeguchi, M.; Komoriya, Y.; Iino, R.; Noji, H. Molecular Mechanism of ATP Hydrolysis in F1-ATPase Revealed by Molecular Simulations and Single-Molecule Observations. *J. Am. Chem. Soc.* **2012**, *134* (20), 8447–8454. https://doi.org/10.1021/ja211027m.

(76) Altschul, S. F.; Gish, W.; Miller, W.; Myers, E. W.; Lipman, D. J. Basic Local Alignment Search Tool. *J. Mol. Biol.* **1990**, *215* (3), 403–410. https://doi.org/10.1016/S0022-2836(05)80360-2.

(77) Altschul, S. F. A Protein Alignment Scoring System Sensitive at All Evolutionary Distances. *J. Mol. Evol.* **1993**, *36* (3), 290–300. https://doi.org/10.1007/BF00160485.

(78) Chakraborty, P.; Hiom, K. DHX9-Dependent Recruitment of BRCA1 to RNA Promotes DNA End Resection in Homologous Recombination. *Nat. Commun.* **2021**, *12* (1), 4126. https://doi.org/10.1038/s41467-021-24341-z.

(79) Melero, R.; Hug, N.; López-Perrote, A.; Yamashita, A.; Cáceres, J. F.; Llorca, O. The RNA Helicase DHX34 Functions as a Scaffold for SMG1-Mediated UPF1 Phosphorylation. *Nat. Commun.* **2016**, *7* (1), 10585. https://doi.org/10.1038/ncomms10585.

(80) Sauer, M.; Juranek, S. A.; Marks, J.; De Magis, A.; Kazemier, H. G.; Hilbig, D.; Benhalevy, D.; Wang, X.; Hafner, M.; Paeschke, K. DHX36 Prevents the Accumulation of Translationally Inactive MRNAs with G4-Structures in Untranslated Regions. *Nat. Commun.* **2019**, *10* (1), 2421. https://doi.org/10.1038/s41467-019-10432-5.

(81) Bosco, B.; Rossi, A.; Rizzotto, D.; Hamadou, M. H.; Bisio, A.; Giorgetta, S.; Perzolli, A.; Bonollo, F.; Gaucherot, A.; Catez, F.; Diaz, J.-J.; Dassi, E.; Inga, A. DHX30 Coordinates Cytoplasmic Translation and Mitochondrial Function Contributing to Cancer Cell Survival. *Cancers (Basel).* **2021**, *13* (17), 4412. https://doi.org/10.3390/cancers13174412.

# Supporting Information

# Molecular Basis of RNA-Driven ATP Hydrolysis in DExH-box Helicases

Santiago Movilla[1], Maite Roca[1], Vicent Moliner[1] and Alessandra Magistrato[2,*]

[1]BioComp Group, Institute of Advanced Materials (INAM), Universitat Jaume I,
12071, Castellón, Spain

[2]Department National Research Council of Italy (CNR) - Institute of Material (IOM)
c/o International School for Advanced Studies (SISSA), Trieste, Italy

[*]Correspondence should be addressed to:

Alessandra Magistrato: alessandra.magistrato@sissa.it

# Table of Contents

## Figures:

## Tables:

## Schemes:

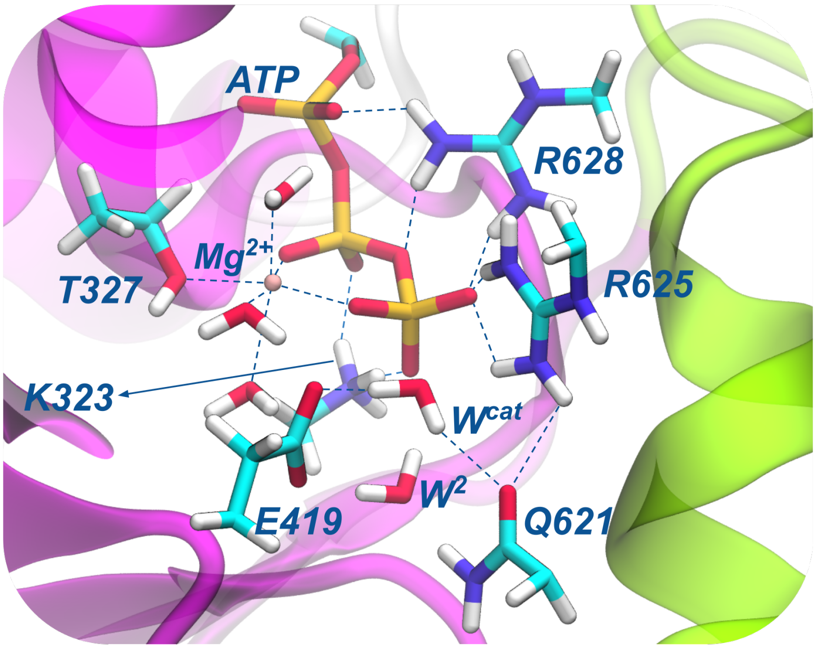


**Figure S1.** General disposition of the Prp2 active site. The H-bond interactions are shown in dashed lines. The region treated quantum mechanically to explore the mechanism of ATP hydrolysis is shown in licorice representation with hydrogen atoms explicitly. The W2 molecule was included in the quantum region after the first step (**R** → **I1**) of the hydrolysis.

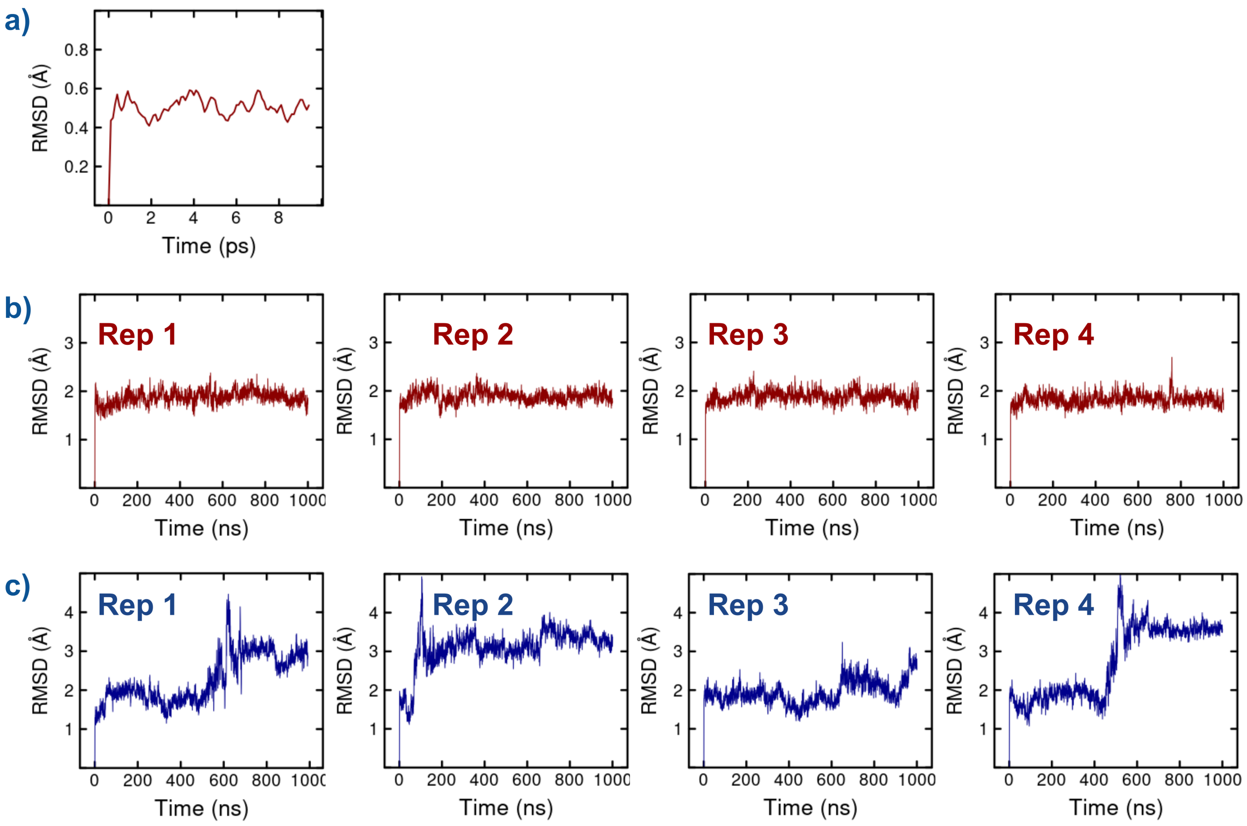


**Figure S2.** Root-mean-square deviation (RMSD) vs. simulation time (ns) of protein backbone atoms (Cα, C, N, O) for the a) QM/MM Molecular Dynamics (MD) simulation in the presence of RNA, and for the four replicas of classical MD simulations b) in presence and c) in the absence of the RNA strand.

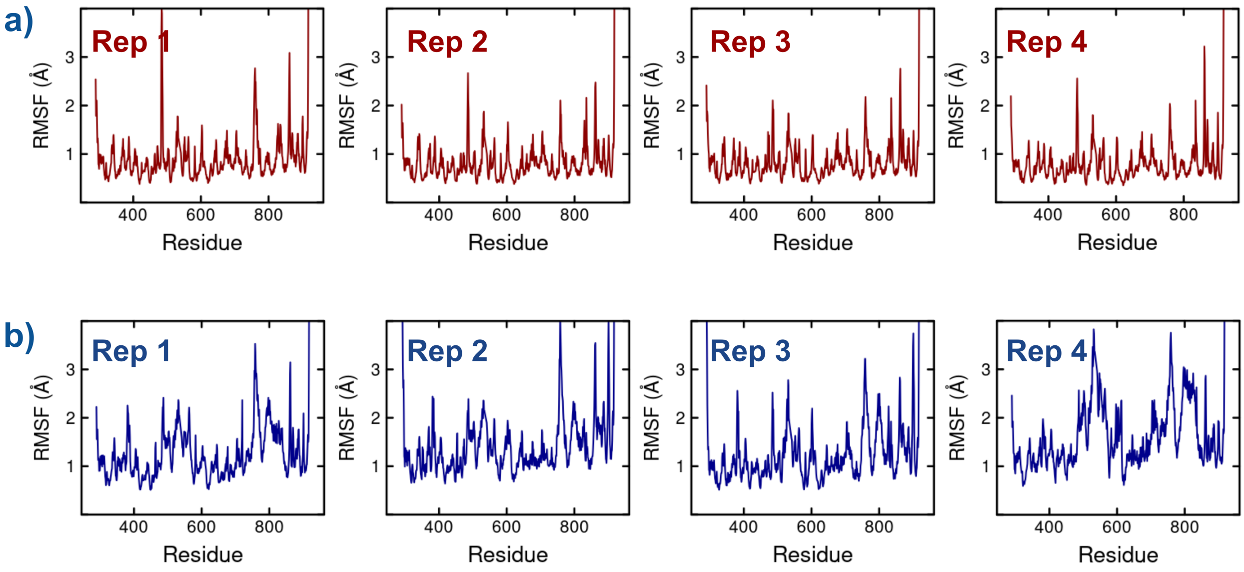


**Figure S3.** Root-mean-square fluctuation (RMSF) analysis of protein Cα atoms using the initial reactive structure as reference for the four replicas of the classical molecular dynamics simulations a) in presence and b) in absence of RNA. Values for some critical residues are explicitly shown in Table S2.

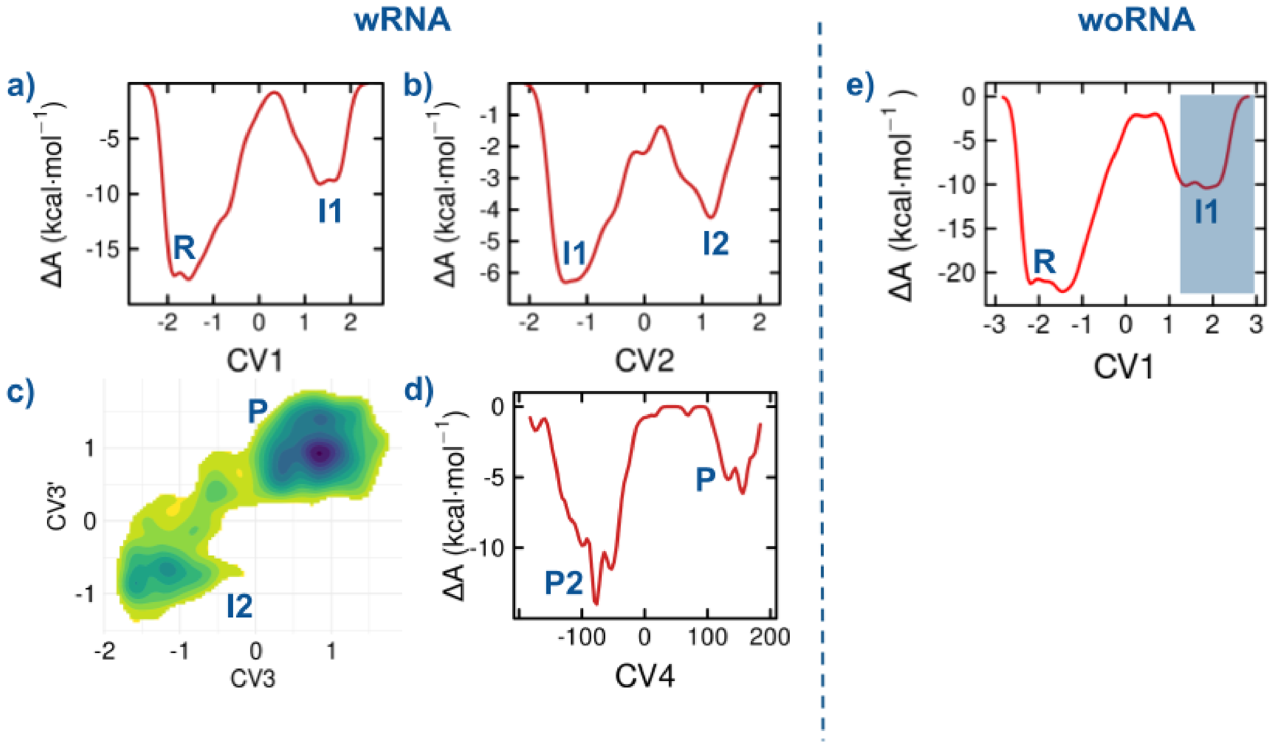


**Figure S4.** (Panels a, b, and d) Helmholtz free energy (ΔA) profile vs. a selected collective variable and c) free energy surface reported for two collective variables for the system with RNA. Panel e) Helmholtz free energy (ΔA) profile vs. a selected collective variable for the system without RNA, the blue rectangle with transparency shows the CV values that were not completely filled during metadynamics. ΔA are computed at DFT-BLYP-D3/MM level for ATP hydrolysis catalyzed by Prp2 and are reported in kcal·mol$^{-1}$. Each level in the contour plot (c) represents a 0.5 kcal·mol$^{-1}$ variation. A complete description of the collective variables and the metadynamics parameters is listed in Table S1.

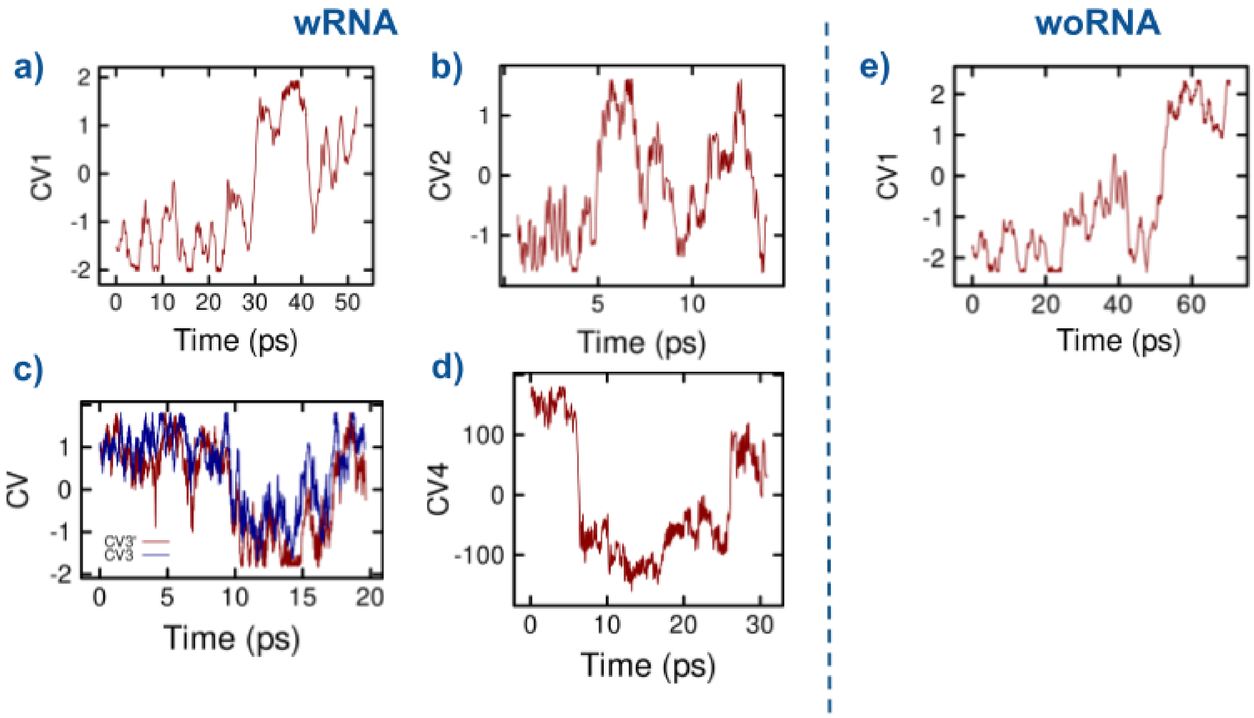


**Figure S5.** Time evolution of the collective variables (CVs) along the metadynamics trajectories (ps) employed to study the ATP hydrolysis mechanism catalyzed by Prp2. Corresponding to a) **R→I1** b) **I1→I2** c) **I2→P** and d) **P →P2** for the system with RNA and e) **R→I1** for the system without RNA. A complete description of the CVs and the metadynamics parameters are listed in Table S1.

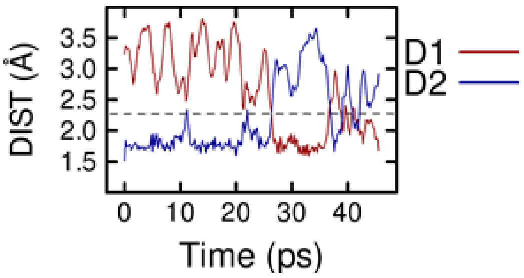


**Figure S6.** Time evolution of D1 ($O_{nuc}$:Wcat -Pγ:ATP) and D2 (Pγ:ATP- Oβ:ATP) distances along the metadynamics (MTD) simulation (ps) of the rate determinant **R→I1** step in the presence of RNA. The approximated transition state distance is shown with a dashed line.

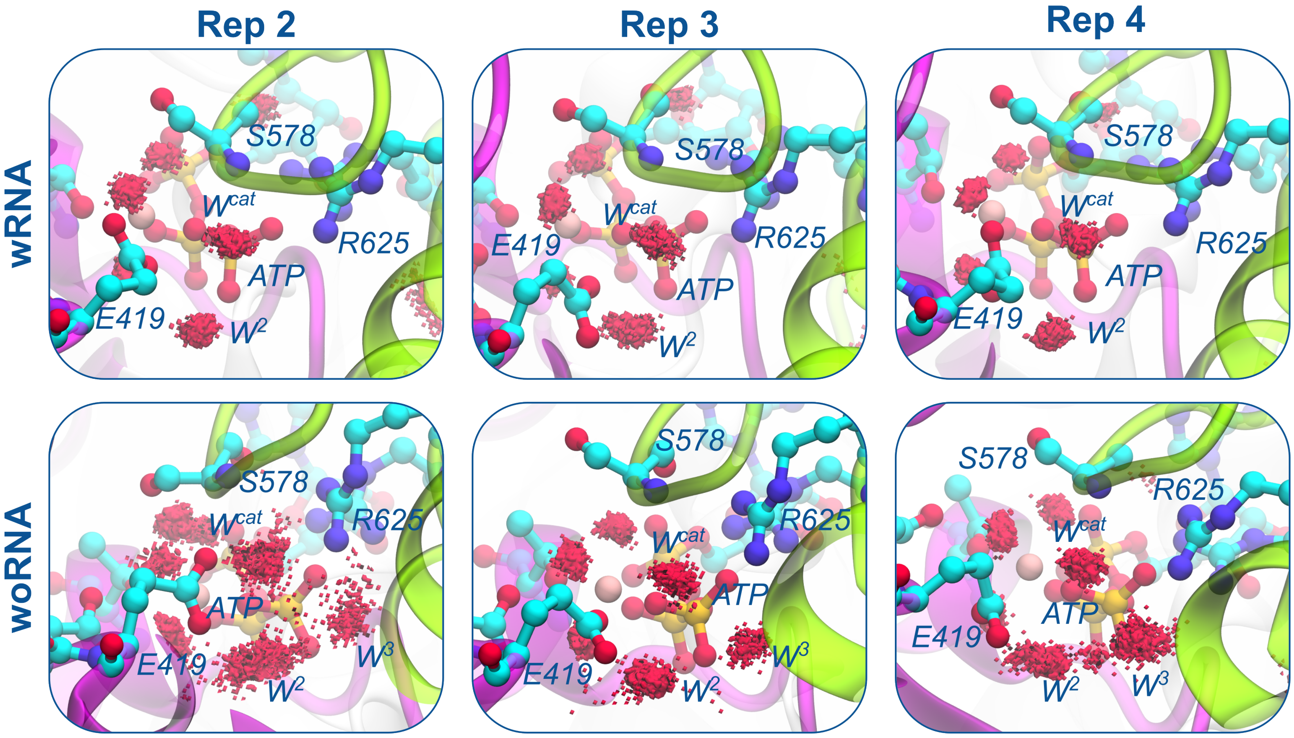


**Figure S7.** Representative frames of the orientation of key Prp2 residues, substrate and density distribution of the positions visited by the nucleophilic and accessory water molecules (Wcat, W2, W3) along the molecular dynamics (MD) simulation trajectories in presence (top panels) and absence (bottom panels) of the 7-nucleotides-long RNA strand for the three 1 µs-long replicas of the classical molecular dynamics simulation.

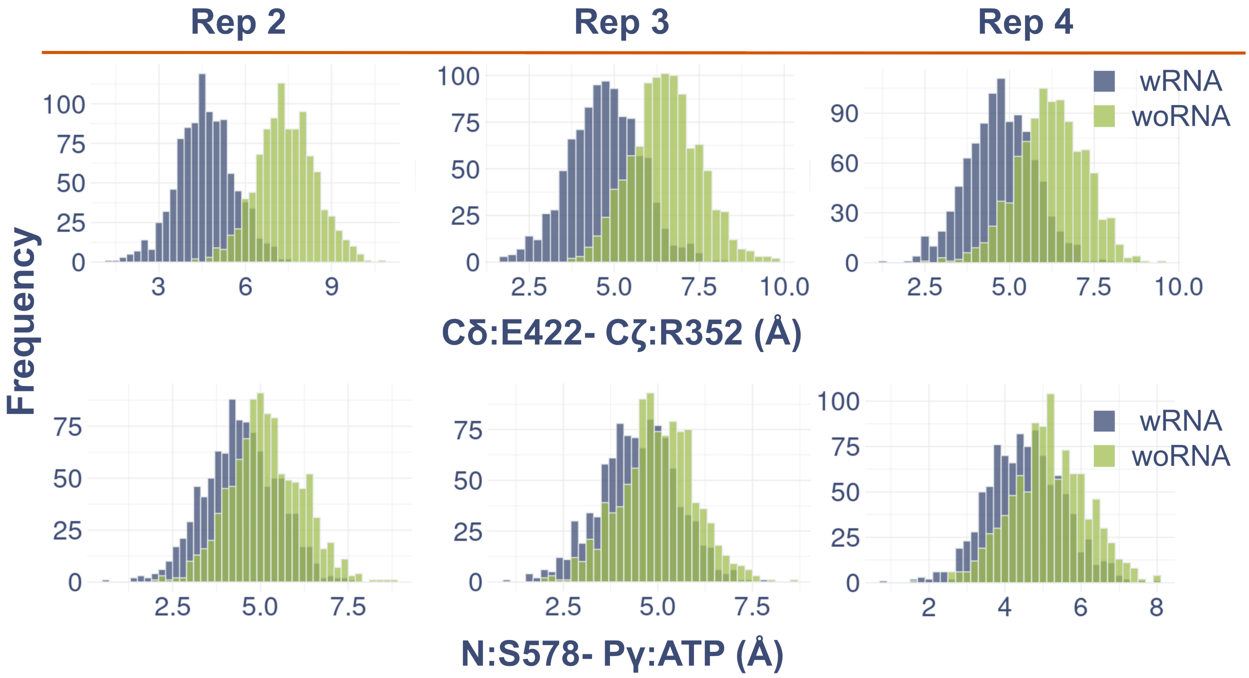


**Figure S8.** Distribution of the distance (Å) between Cδ:Glu419-Cζ:Arg352 atoms (top panels) for the three additional 1 µs-long classical molecular dynamics (MD) simulations replicas and distribution of the distance (Å) between N:Ser578-Pγ:ATP atoms (bottom panels) along the three 1 µs-long classical MD replicas in the presence (blue) and absence (green) of the RNA strand.

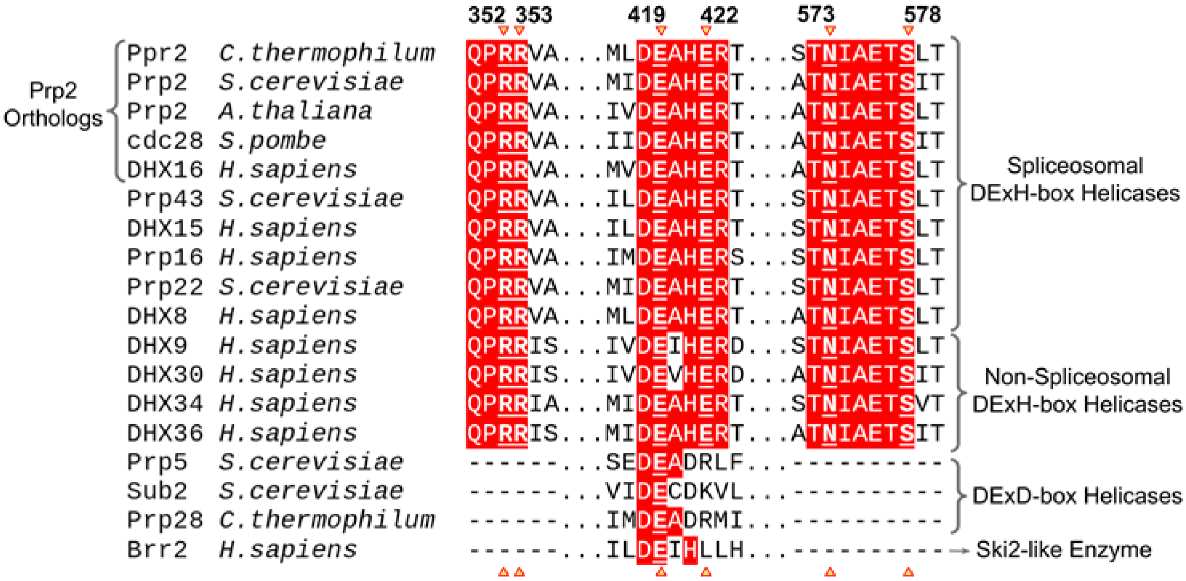


**Figure S9.** Sequence alignment among Prp2, various DExH-box ATPases/helicases, spliceosomal DExD-box ATPases/helicases and Brr2 helicase of the Ski-like family. The conserved residues are shaded red and those involved in allosteric modulation are marked in bold and underlined.

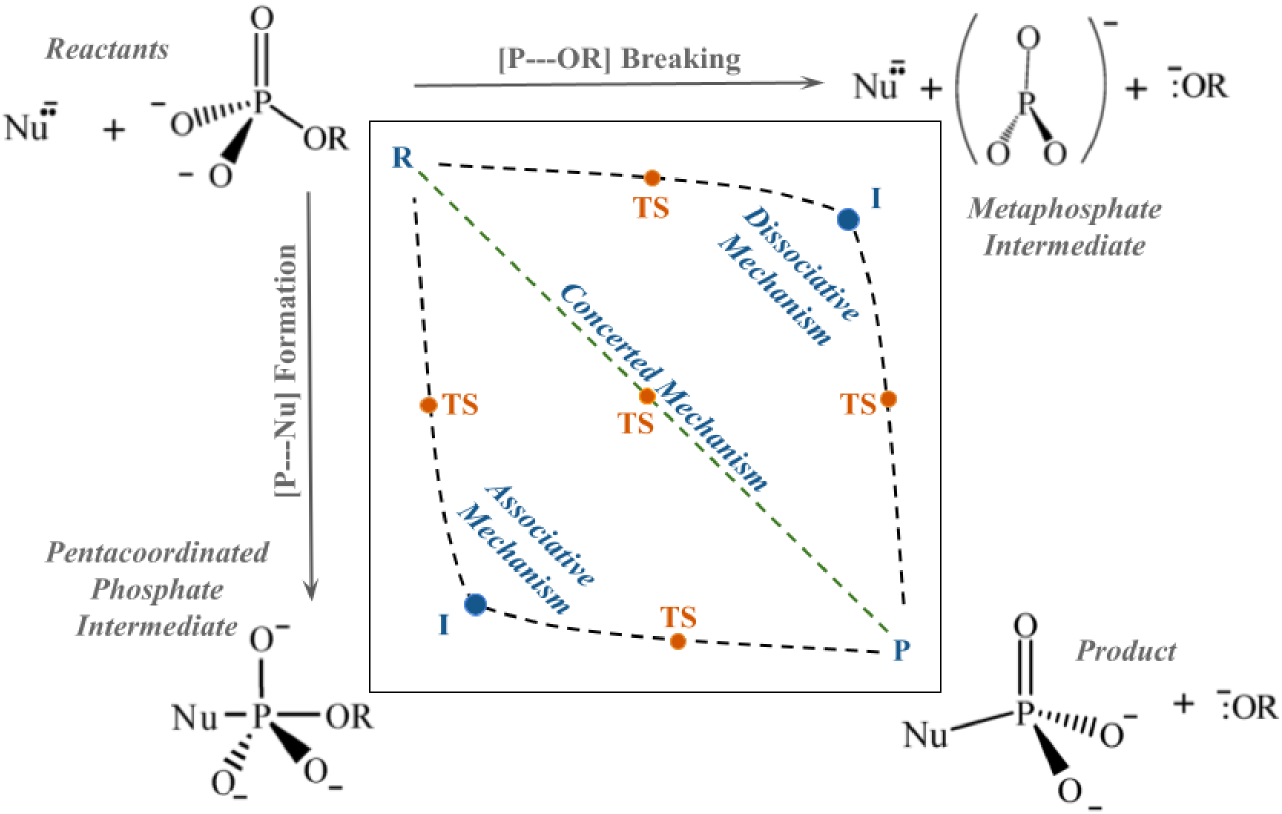
**Table S1.** Detailed description of the collective variables (CVs) and parameters of the metadynamics simulations used to compute the reaction mechanism catalyzed by the Prp2 ATPase/helicase.

| CV | Coordinate | Width | MD Time Step (fs) | Time Between Hills (fs) |
|---|---|---|---|---|
| 1 | [$\mathrm{O_{nuc}}$:Wcat-Pγ:ATP]-[Pγ:ATP- Oβ:ATP] | 0.158 Å (0.3 a.u.) | 0.25 | 30 |
| 2 | [HOε:Glu419-$\mathrm{O_{nuc}}$:Wcat]-[HOε:Glu419-O:W2] | 0.324 Å (0.625 a.u.) | 0.5 | 30 |
| 3 | [HOε:Glu419-Oε:Glu419]-[HOε:Glu419-O:W2] | 0.106 Å (0.2 a.u.) | 0.25 | 30 |
| 3’ | [H:W2-O:W2]-[H:W2-O:$\mathrm{HPO_4^{2-}}$] | 0.106 Å (0.2 a.u.) | 0.25 | 30 |
| 4 | HO$\gamma_1$-O$\gamma_1$-Pγ-O$\gamma_2$:$\mathrm{H_2PO_4^-}$ | 0.09 rad | 0.5 | 30 |

**Table S2.** Average values of the root-mean-square fluctuation (RMSF) of Cα atoms using the initial reactive structure as reference for the residues involved in the RNA-driven allosteric activation of the Prp2 ATPase/helicase function along the four classical MD replicas.

| **Resid** | **RMSF wRNA (Å)** | **RMSF woRNA (Å)** |
|---|---|---|
| Arg352 | 0.51±0.07 | 1.10±0.26 |
| Arg353 | 0.48±0.04 | 1.16±0.21 |
| Glu419 | 0.39±0.01 | 0.67±0.18 |
| Glu422 | 0.46±0.04 | 0.87±0.23 |
| Asn573 | 0.47±0.03 | 1.08±0.29 |
| Ser578 | 0.44±0.03 | 0.71±0.16 |

**Scheme S1.** More-O'Ferrall–Jencks representation of the general plausible reaction pathways for phosphate lysis. Three possible mechanisms: (i) the dissociative mechanism proceeds through an $S_N1$ in which the leaving group is dissociated from the phosphorus atom. This generates a metaphosphate intermediate, which is then attacked by the nucleophile (Nu, in our case a water molecule), ii) the associative mechanism proceeds through an intermediate with a pentacoordinated phosphorus atom that subsequently liberates the leaving group to give way to the products and iii) the concerted mechanism proceeds through a single transition state in which the bonds with the leaving group and with the nucleophile are concertedly broken and formed, respectively. Notably, the concerted mechanism can proceed synchronously or asynchronously, but always in a single continuous transition. In the diagram the minima are indicated in blue while the transition states are highlighted in orange. The dotted lines describe the reaction path associated with each mechanism as described by the evolution of the P-OR (horizontal axis) and P-Nu (vertical axis) distances. The reaction pathway by which Prp2-catalyzed ATP hydrolysis proceeds is indicated in green.



# ANNEX D. CO-AUTHOR'S CONSENT

UNIVERSITAT
JAUME I
Escola de Doctorat · ED

Castelló de la Plana, 10h November 2022

I, Sergio Martí, hereby authorise Santiago Movilla Núñez to include the list of publications listed below in his doctoral thesis. In addition, I waive the right to use those articles or/and manuscripts submitted for publication as part of any other doctoral thesis.

List of articles or/and manuscripts submitted for publication:

(1) Movilla, S.; Martí, S.; Roca, M.; Moliner, V. Unrevealing the Proteolytic Activity of RgpB Gingipain from Computational Simulations. *J. Chem. Inf. Model.* **2021**, 61 (9), 4582–4593. https://doi.org/10.1021/acs.jcim.1c00666.

(2) Movilla, S.; Martí, S.; Roca, M.; Moliner, V. Computational Study of the Inhibition of RgpB Gingipain, a Promising Target for the Treatment of Alzheimer's Disease. *Submitted.*

Signed,

SERGIO|
MARTI|
FORES

Firmado digitalmente
por SERGIO|MARTI|
FORES
Fecha: 2022.11.10
08:57:17 +01'00'

*In accordance with article 28 of the regulations on doctoral studies of the Universitat Jaume I in Castelló, regulated by RD 99/2011, at the Universitat Jaume I (Approved by the Governing Council at its meeting no. 8/2020 held on 2th October 2020):*
*"(...)*
*4. In the case of joint publications, all the co-authors must explicitly state their approval that the doctoral student presented the work as part of her/his thesis and the express waiver of presenting this same work as part of another doctoral thesis. This authorisation must be attached as documentation when the evaluation of the thesis begins"*

UNIVERSITAT JAUME I
Escola de Doctorat · ED

Castelló de la Plana, 10th November 2022

I, María Teresa Roca, hereby authorise Santiago Movilla Núñez to include the list of publications listed below in his doctoral thesis. In addition, I waive the right to use those articles or/and manuscripts submitted for publication as part of any other doctoral thesis.

List of articles or/and manuscripts submitted for publication:

(1) Movilla, S.; Martí, S.; Roca, M.; Moliner, V. Unrevealing the Proteolytic Activity of RgpB Gingipain from Computational Simulations. *J. Chem. Inf. Model.* **2021**, 61 (9), 4582–4593. https://doi.org/10.1021/acs.jcim.1c00666.

(2) Movilla, S.; Martí, S.; Roca, M.; Moliner, V. Computational Study of the Inhibition of RgpB Gingipain, a Promising Target for the Treatment of Alzheimer's Disease. *Submitted.*

(3) Movilla, S.; Roca, M.; Moliner V.; Magistrato, A. Molecular Basis of RNA-Driven ATP Hydrolysis in DExH-box Helicases. *Submitted.*

Signed,

MARIA
TERESA|
ROCA|
MOLINER

Firmado
digitalmente por
MARIA TERESA|
ROCA|MOLINER
Fecha: 2022.11.10
09:16:07 +01'00'

*In accordance with article 28 of the regulations on doctoral studies of the Universitat Jaume I in Castelló, regulated by RD 99/2011, at the Universitat Jaume I (Approved by the Governing Council at its meeting no. 8/2020 held on 2th October 2020):*
*"(...)*
*4. In the case of joint publications, all the co-authors must explicitly state their approval that the doctoral student presented the work as part of her/his thesis and the express waiver of presenting this same work as part of another doctoral thesis. This authorisation must be attached as documentation when the evaluation of the thesis begins"*

UNIVERSITAT
JAUME I
Escola de Doctorat · ED

Castelló de la Plana, 10th November 2022

I, Vicent Moliner, hereby authorise Santiago Movilla Núñez to include the list of publications listed below in his doctoral thesis. In addition, I waive the right to use those articles or/and manuscripts submitted for publication as part of any other doctoral thesis.

List of articles or/and manuscripts submitted for publication:

(1) Movilla, S.; Martí, S.; Roca, M.; Moliner, V. Unrevealing the Proteolytic Activity of RgpB Gingipain from Computational Simulations. *J. Chem. Inf. Model.* **2021**, 61 (9), 4582–4593. https://doi.org/10.1021/acs.jcim.1c00666.

(2) Movilla, S.; Martí, S.; Roca, M.; Moliner, V. Computational Study of the Inhibition of RgpB Gingipain, a Promising Target for the Treatment of Alzheimer's Disease. *Submitted.*

(3) Movilla, S.; Roca, M.; Moliner V.; Magistrato, A. Molecular Basis of RNA-Driven ATP Hydrolysis in DExH-box Helicases. *Submitted.*

Signed,

VICENTE
MOLINER IBAÑEZ
- NIF:25388050Y

Firmado digitalmente por
VICENTE MOLINER IBAÑEZ -
NIF:25388050Y
Fecha: 2022.11.10 08:44:15
+01'00'

*In accordance with article 28 of the regulations on doctoral studies of the Universitat Jaume I in Castelló, regulated by RD 99/2011, at the Universitat Jaume I (Approved by the Governing Council at its meeting no. 8/2020 held on 2th October 2020):*

*"(...)*

*4. In the case of joint publications, all the co-authors must explicitly state their approval that the doctoral student presented the work as part of her/his thesis and the express waiver of presenting this same work as part of another doctoral thesis. This authorisation must be attached as documentation when the evaluation of the thesis begins"*

UNIVERSITAT
JAUME I
Escola de Doctorat · ED

Trieste, 10th November 2022

I, Alessandra Magistrato, hereby authorise Santiago Movilla Núñez to include the list of publications listed below in his doctoral thesis. In addition, I waive the right to use those articles or/and manuscripts submitted for publication as part of any other doctoral thesis.

List of articles or/and manuscripts submitted for publication:

(1) Movilla, S.; Roca, M.; Moliner V.; Magistrato, A. Molecular Basis of RNA-Driven ATP Hydrolysis in DExH-box Helicases. *Submitted.*

Signed,

Alessandra
Magistrato
10.11.2022
09:36:16
GMT+01:00

*In accordance with article 28 of the regulations on doctoral studies of the Universitat Jaume I in Castelló, regulated by RD 99/2011, at the Universitat Jaume I (Approved by the Governing Council at its meeting no. 8/2020 held on 2th October 2020):*

*"(...)*

*4. In the case of joint publications, all the co-authors must explicitly state their approval that the doctoral student presented the work as part of her/his thesis and the express waiver of presenting this same work as part of another doctoral thesis. This authorisation must be attached as documentation when the evaluation of the thesis begins"*